Supplementary Information

An early modern human from Romania with a recent Neanderthal ancestor

Supplementary Information 1

Mitochondrial DNA and filtering of contaminated sequences

Mitochondrial contamination

We captured mitochondrial DNA genome sequences (mtDNA) from all five libraries of the Oase 1 individual using the in-solution capture method of [1].

We analyzed a total of 27,958,007 sequences that all had perfect matches to the expected indices (Extended Data Table 6). We merged the reads from either end of the same molecule by requiring an overlap of at least 11 base pairs, and generated a consensus sequence by picking the base with the higher sequence quality[2]. We mapped these merged reads, which we call a "fragment", to the revised Cambridge reference sequence (rCRS NC_012920). We restricted analysis to Oase 1 fragments that were at least 35bp and that had a mapping quality of at least 30. We removed duplicate fragments by identifying all fragments with the same start and stop positions and keeping the one with the highest average base quality.

The deamination of cytosine (C) to uracil (U) residues, which occurs primarily at single-stranded DNA overhangs, leaves characteristic patterns of C→T substitutions in sequences obtained from ancient DNA molecules, because uracils are read as thymines (T) by DNA polymerases. We measured the frequency of C→T substitutions relative to the mtDNA consensus sequence, which we obtained as described below. The frequency at the 5' end was 8% and 7% for UDG-treated and 20%, 21% and 20% for non-UDG treated libraries. The frequency at the 3' end was 19% and 25% for UDG-treated and 21%, 20% and 20% for non-UDG treated libraries (Extended Data Table 6). The C→T rates at the 5' end increased when we restricted analysis to fragments that carry a C→T substitution at the opposite end of the same fragment, consistent with a mixture of contamination and endogenous ancient DNA[3].

We obtained direct evidence for contamination by examining the consensus of all Oase 1 fragments, and the consensus of fragments that contained C→T substitutions at the terminal ends. When all fragments are analyzed, the consensus corresponds to a derived haplogroup (H39) of macrohaplogroup N, which is widespread in present-day non-Africans, especially in West Eurasia. In contrast, when we restrict to fragments that have C→T substitutions at the terminal ends, the consensus does not have any of the mutations that occurred on the lineage leading to haplogroup H39 since the inferred ancestor of all copies of macrohaplogroup N observed to date (Extended Data Figure 1). We aligned the Oase 1 consensus mtDNA sequence from all fragments to mtDNA sequences from 10 other securely dated archaeological samples[3-6], as well as to mtDNA sequences from 311 present-day humans. Based on the number of mutations missing relative to the most closely related mtDNA sequences, as determined by running the MrBayes software[7] on the joint dataset using the same procedure described in[6], we estimate the date of the Oase1 mtDNA consensus to be 7,111 years before present (95% highest posterior density 96-13,735 before present), consistent with contamination from a present-day human.

To estimate the proportion of mtDNA contamination, we determined positions in the mtDNA that are specific ('diagnostic') for Oase 1. This required generating a consensus mtDNA sequence for Oase 1. To generate this consensus, we restricted to the subset of Oase 1 fragments that passed the filters in Box S1.1. These are the same filters that we apply to the nuclear data and that we use for population genetic analysis. For the fragments that passed

these filters, we masked nucleotides in the three final positions in the same orientation as sequenced, as they are prone to have arisen from cytosine deamination.

Box S1.1: Filters used to restrict to deaminated fragments

UDG-treated libraries: Restrict to fragments with C→T substitutions in the first position at the 5'- and the last two positions on the 3'-end. These are the only bases that largely escape uracil removal using the protocol we used to build these libraries [8].

Non-UDG-treated libraries: Restrict to fragments with C→T substitutions in the first three positions at the 5'- and the last three positions bases on the 3'-end. The nucleotides at the ends of molecules are the ones that most likely to harbor uracils.

To determine a consensus base at each position of the mitochondrial genome, we required a minimum coverage of 5 and a consensus support of ≥80% of fragments. We obtained unambiguous base call for all but 5 positions (Table S1.1). Positions 297 and 310 are in the C-homopolymer stretch, which is known to be a problematic region for mtDNA alignment. Positions 514 and 515 are also in a short repetitive sequence. Position 16293 has 69% support for the majority nucleotide (G), but when we restrict to the fragments sequenced on the forward strand where G nucleotides are not vulnerable to ancient DNA degradation, the support is 97% (only 1 of 29 fragments disagreeing), giving us confidence that G is the true base. We do not use the four ambiguous positions 297, 310, 514 and 515 in mtDNA analysis.

Table S1.1. Positions with support from <80% of fragments

rCRS	Base in the	Majority	Called	Coverage	Fraction of fragments	Used in mtDNA
297	A	G	N	85	73%	No
310	T	C	N	87	60%	No
514	C	C	N	42	62%	No
515	A	A	N	45	60%	No
16293	A	G	N	45	69%	Yes

To estimate a contamination rate, we identified 6 positions where the consensus of Oase 1 differs from at least 99% of a panel of 311 present-day human mtDNAs[9] (Table S1.2).

Table S1.2. Diagnostic positions for Oase 1. There are 6 positions where the consensus mtDNA of Oase 1 differs from at least 99% of 311 present-day human mtDNA genomes

rCRS position	Oase 1	Consensus	Frequency of allele
3205	A	C	100%
3462	T	C	100%
4232	C	T	≥99%
7158	G	A	≥99%
8749	C	T	≥99%
11016	A	G	≥99%

We counted the sequences that overlap these positions to determine the fraction mismatching Oase 1, taking into account the strand orientation in the cases where the informative sites are C or G. If the informative state is C, we counted only the alignments on the reverse strand, and in cases where the informative state is G, we counted only the alignments on the forward strand. Before filtering, the estimates of contamination for all five libraries range from 59-73% and combining the data from all five libraries the estimate is 67% (95% confidence interval 65% to 69%) (Table S1.3). When we applied the filters in Box S1.1, the estimates of contamination for all 5 libraries range from 0% to 7% and combining the data from all five libraries the estimate is 4% (95% CI 2% to 9%) (Table S1.3).

Table S1.3. Estimates of contamination in Oase 1. We estimate contamination based on the rate of mismatch to the consensus at six sites that are diagnostic for Oase 1.

Library ID	#fragments	Coverage	Contamination (fraction of observations)
No filtering			
A5227	34,486	112	59% (172/193)
A5252	31,368	114	64% (197/310)
A9032	51,810	178	69% (316/460)
A9033	55,878	193	70% (341/485)
A9034	59,883	207	70% (365/523)
2 UDG-treated	65,854	226	61% (369/603)
3 non-UDG-treated	167,571	578	70% (1022/1468)
5 libraries together	233,425	804	67% (1391/2071)
C→T filtering			
A5227	1,784	4.9	0% (0/16)
A5252	1,569	4.6	0% (0/11)
A9032	6,612	20.4	7% (2/30)
A9033	7,171	22.1	5% (3/55)
A9034	7,627	23.5	4% (2/45)
2 UDG-treated	3,353	9.5	0% (0/31)
3 non-UDG-treated	21,410	66.0	5% (7/130)
5 libraries together	24,763	75.5	4% (7/161)

Relationship of the Oase 1 mtDNA to that of present-day humans

We identified the haplogroup for Oase 1 from deaminated fragments using HaploGrep[10] based on the Phylotree database (Phylotree.org, build 16). Oase 1 carries the following substitutions that define the N macrohaplogroup:

73G, 263G, 750G, 1438G, 2706G, 3107d, 4769G, 7028T, 8860G, 11719A, 12705T, 14766T, 15326G, 16223T

Oase 1 does not share derived alleles at positions 8701 (G in Oase 1) and 9540 (C in Oase 1) that have been observed in all copies of macrohaplogroup N observed to date. This suggests that the consensus of the Oase 1 deaminated fragments derives from a lineage that diverged from the stem of macrohaplogroup N and that has never previously been sampled.

To generate a phylogenetic tree, we added the consensus sequences of all fragments and deaminated fragments from Oase 1 to the mtDNA sequences of three early modern humans (Ust'-Ishim, Kostenki 14, and Tianyuan), 54 present-day humans[11] and a Neanderthal mtDNA (Vindija 33.25)[12]. We applied the software MrBayes[7] and ran 20,000,000 iterations of the Markov Chain Monte Carlo with the first 2,000,000 iterations discarded as burn-in. We used a General Time Reversible sequence evolution model with a fraction of invariable sites (GTR+I) determined by the best-fit model approach of Modeltest and PAUP*[13]. Extended Data Figure 1 shows the resulting mitochondrial DNA tree. The consensus of all fragments clusters with most present-day Europeans (100% posterior support). After restricting to deaminated fragments, the consensus of Oase 1 belongs to macrohaplogroup N, but it branches off before all copies of this macrohaplogroup observed to date (100% posterior support), consistent with the expectation for a very ancient sample.

We estimated the date of the Oase 1 mtDNA using BEAST[14] by co-analyzing it with the mtDNAs of 311 present-day humans and ten securely radiocarbon dated ancient humans[4,5]. These are all the samples analyzed in [4] to which we added Ust'-Ishim[6]. We carried out two Markov Chain Monte Carlo (MCMC) runs with 30,000,000 iterations each, sampling every 1000 steps and using both constant and strict clock models. The first 6,000,000 iterations were discarded as burn-in. For each model, both independent runs were combined, resulting in 48,000,000 iterations. We estimate that the Oase 1 mandible dates to 36,328 BP (95%

highest posterior density of 14,515-56,452 BP). Although the confidence interval is large due to the small size of the mtDNA genome, it is consistent with the radiocarbon date of 37,615-41,761 BP and provides further evidence for the authenticity of the ancient sequences.

Nuclear contamination

We analyzed the sequences mapping to the 2,051,902 unique SNP targets from nuclear capture Panels 1, 2 and 3. The number of SNPs covered at least once by any Oase 1 fragment was 1,038,619 (union of all SNPs). Of these, 271,326 were covered by at least one deaminated fragment based on the criteria in Box S1.1.

To test for evidence of contamination, we computed D-statistics using ADMIXTOOLS [15] on the intersection of SNPs covered by both deaminated and non-deaminated fragments: *D(Oase1-deaminated, Oase1-non-deaminated; Test, Africa)*. Here, Africa is represented by a pool of six genomes (2 Yoruba, 2 Mbuti, and 2 Dinka). We find that when *Test* is European (represented by a pool of 2 French and 2 Sardinian genomes), *Test* shares significantly more alleles with Oase 1 when all fragments are analyzed than when only deaminated fragments are analyzed (Z = -18.8). In contrast, when *Test* is East Asian (represented by a pool of 2 Han and 2 Hai genomes), the skew is less (Z = -5.2). This is consistent with Oase 1 being contaminated by DNA that is more European than East Asian.

To confirm that our filtering for deaminated fragments reduces the impact of contamination not just for mtDNA but also for nuclear data, we computed statistics of the form *D(Loschbour, East Asia; Oase 1, Africa)*, representing Oase 1 alternately by all fragments or just by deaminated fragments. Loschbour is an ~8,000 year old Mesolithic hunter-gatherer from Luxembourg in Western Europe [16]; East Asia is a pool of four genomes (2 Han and 2 Dai); and Africa is a pool of six genomes (2 Yoruba, 2 Mbuti and 2 Dinka). We do not use present-day Europeans to represent Europeans, and instead use a Mesolithic European, because present-day Europeans (but not Mesolithic Europeans) have evidence of ancestry from a population that split from present-day eastern and western non-Africans before they separated from each other [16]. This would be expected to bias the *D*-statistic negative even in the absence of contamination, making it difficult to interpret evidence of contamination.

Table S1.4 shows that there is a significantly positive *D* when all Oase 1 fragments are used, consistent with European contamination (95% CI 0.0029 to 0.0046). When we restrict to deaminated Oase 1 fragments, *D* is diminished, with a confidence interval that does not overlap all fragments (95% CI -0.0014 to 0.0008). Thus, restricting to deaminated fragments reduces contamination, and may effectively eliminate it since the *D* range overlaps zero.

Table S1.4. Statistics of the form *D*(*Loschbour, East Asian; Oase 1, African)*

Fragments used	Sites	D (Estimate)	D (Std. Err.)	D (95% CI)	Z-score (deviation from 0)
All	997,700	0.0037	0.0004	(0.0029, 0.0046)	8.9
Deaminated only	261,947	-0.0003	0.0006	(-0.0014, 0.0008)	-0.5

Under the assumption that the contaminating DNA is entirely from one or more European individuals, and that the deaminated fragments are uncontaminated, we can estimate the proportion of nuclear contamination by modeling all the Oase 1 fragments as a mixture of present-day European and uncontaminated Oase 1. Mathematically, this is the same as the problem of estimating the proportion of European mixture in the Oase 1 all fragments dataset, given data from two reference populations that we propose to be sister groups to present-day Europeans on the one hand, and deaminated Oase 1 fragments on the other hand. This problem has been addressed in the literature on estimating mixture proportions, and we

borrow that technology. Methodological details are given in[17].

We estimated nuclear contamination on a merge of Human Origins genotyping data reported in [16] with the Oase 1 data, as we need data from more populations than are available from sequencing data. We divide the populations into two sets. The *left* set *L* consists of the proposed admixed population (Oase 1 all fragments) and the putative clades with the source populations (Europeans, and Oase 1 deaminated fragments). The *right* set *R* consists of 15 worldwide populations excluding Europeans (Ami, Biaka, Bougainville, Chukchi, Eskimo, Han, Ju_hoan_North, Karitiana, Kharia, Mbuti, Onge, Papuan, She, Ulchi, Yoruba); this is the same set of *right* populations used in[17]. The *right* populations are variably related to West Eurasians, which provides leverage for discerning different components of ancestry among the *left* populations.

We estimate the matrix of f_4-statistics $M(l, r) = f_4(l_x, l; r_x, r)$ where l_x and r_x are fixed reference populations in *L* and *R* respectively, and *l, r* are other populations in *L* and *R*. We of course cannot know the true matrix, but can estimate it from the observed f_4-statistics, and we can also estimate a covariance matrix using a Block Jackknife. If the three *left* populations are related to the *right* populations via just two ancestral lineages (from which Oase 1 deaminated fragments and the Europeans directly descend), and if all Oase 1 fragments can be modeled as a mixture of these two lineages, then the matrix should be Rank 1 (using the terminology of linear algebra). In contrast, if the Oase 1 deaminated fragments are uncontaminated, we expect the matrix to be Rank 0.

Applying this test to our data, we find that we can reject the hypothesis that the matrix is Rank 0 ($P = 1.0\times10^{-9}$). There is weak evidence that we can also reject Rank 1 (P=0.014), suggesting that most but not all of the contamination is coming from Europeans.

The evidence of a rejection of Rank 1 is modest (European contamination appears to explain most of the rejection from Rank 0), and we therefore attempted to estimate contamination under a Rank 1 model. Intuitively, if *M(l,r)* is Rank 1, Oase 1 fragments are a mixture of just two lineages related to Oase 1 deaminated fragments and to Europeans. In this case, the f_4-statistics relating Oase 1 all fragments to other populations will be a linear combination of the f_4-statistics relating these source populations to the other samples. Thus, we can empirically learn the weights and interpret these as mixture proportions. When we apply this procedure, using the implementation that is reported in[17], we estimate 23.3% ± 3.4% European-related contamination in Oase 1. The implied estimate of nuclear contamination of around 16.7%-29.9% (95% CI) is substantially less than the estimate of around two thirds for mitochondrial DNA fragments (Table S1.3). This is not necessarily a contradiction, however, as the ratio of mitochondrial to nuclear sequences is known to fluctuate in ancient DNA libraries[6,18]. There is no *a priori* reason to expect that the contamination rate should be the same for these two compartments of the genome.

Supplementary Information 2

The Y chromosome of Oase 1

Data processing and sex determination
Panels 1-3 included targets on both chromosome X and Y, so we could use these fragments to determine the sex of Oase 1. Because of the evidence of contamination documented in Supplementary Information section 1, we restricted to deaminated fragments (Methods).

We compared the number of SNPs matching to chromosome X and Y targets in Oase 1, to the number of targeted SNPs. Table S2.1 shows that the fraction of SNPs captured is similar for chromosome X (6.2%) and chromosome Y (8.6%). We conclude that the individual is a male since we do not expect to identify sequences from the Y-chromosome if the individual is female. This is consistent with the morphology[19].

Table S2.1. Number of SNPs covered at least once on chromosome X and Y.

	Chromosome X	Chromosome Y
SNPs hit at least once with a deaminated fragment	3446	2829
Targeted SNPs	55343	32768
Fraction of targeted SNPs hit	6.2%	8.6%

Y-chromosome haplogroup
We determined the Y-haplogroup based on the ISOGG database version 10.14, which gives haplogroup assignments for a subset of SNPs on the Y chromosome (http://www.isogg.org/tree). There are 754 SNPs (out of the 15,102 in this version of the ISOGG database) that are covered at least once in Oase 1. We used these to determine the position of the Oase 1 Y chromosome in the tree based on where it carried the derived or ancestral allele. This allowed us to define Oase 1 as belonging to macrohaplogroup F (positions given in *hg19* coordinates).

Assignment to F:
P187 (9108252 G→T); P158 (17493513 C→T)

Assignment to CF (which contains CT):
CTS6376 (16863259 C→G)

Assignment to CT (which contains F and CF):
PF38 (3396403 C→T); M5612 (7782393 C→T); M5631 (8396636 G→A); M5632 (8526565 G→A); Z17706 (9989244 G→T); Y1525 (14074463 C→T); L957 (14079528 C→T); Y1526 (14472971 C→T); CTS3662 (15097073 G→A); CTS8542 (18077583 T→C); M5760 (18974195 C→T); L1480 (19212465 A→G); M5783 (21429988 A→G); M5786 (21650381 A→G); Z17721 (22477665 G→C); M5809 (23090404 G→A); M5812 (23105586 C→A); M5823 (23567930 C→T)

We found no evidence that Oase 1 belongs to specific sub-haplogroups of F, as it carries the ancestral allele at all previously described diagnostic mutations for these haplogroups.

However, we cannot rule out the possibility that Oase 1 is derived at other SNPs that are diagnostic for these sub-haplogroups but for which Oase 1 has no coverage.

No evidence of membership in G:
S8863 (4179056: G→A); M3485 (8563874 C→T); M3486 (8600158 A→T); L154 (8614138 T→G); M3496 (9850420 C→A); M3497 (9850423 C→A); Z3248 (13460729 G→A); CTS5317 (16203361 G→C); Z3400 (18744995 T→C); M3569 (18744996 C→T); PF3083 (22272581 T→C); PF3087 (22472842 A→C); CTS10945 (22848965 A→G)

No evidence of membership in H:
Z13965 (24523481 C→G)

No evidence of membership in IJ:
P127 (8590752 C→T); PF3526 (8590752 C→T)

We could not test whether Oase 1 is part of macrohaplogroups GHIJK, HIJK, K or K(xLT), as none of the SNPs diagnostic for them are covered by deaminated fragments in Oase 1.

Supplementary Information 3

Relationship of Oase 1 to other genome sequences

Oase 1 is more closely related to non-Africans than to Africans

After restricting to deaminated fragments as described in Supplementary Information section 1, we computed *D*-statistics, restricting to SNPs in Panel 1-3. We show statistics both for all sites that pass the filters, and also restricting to transversions to mitigate against the possibility that ancient DNA degradation is biasing our results. Our findings from the two classes of sites are qualitatively consistent.

We first tested whether Oase 1 shares more alleles with selected African or non-African individuals using the statistic *D(African, Non-African; Oase 1, Chimp)*. Table S3.1 shows that Oase 1 is more closely related to non-Africans: Z << -22 standard errors below 0.

Table S3.1. *D*-statistics of the form *D(African, non-African; Oase 1, Chimp).*

		Transversions only			All sites		
African	**Non-African**	**D**	**Z-score**	**Sites used**	**D**	**Z-score**	**Sites used**
$Mbuti_B$	Ust'-Ishim	-0.037	-36.1	111,996	-0.038	-40.9	254,933
$Mbuti_B$	Kostenki14	-0.039	-34.7	105,160	-0.039	-42.9	240,453
$Mbuti_B$	MA1	-0.039	-32.5	81,016	-0.040	-42.1	186,007
$Mbuti_B$	Loschbour	-0.037	-37.0	111,207	-0.040	-45.8	252,665
$Mbuti_B$	LaBrana	-0.038	-34.5	107,976	-0.041	-42.7	246,895
$Mbuti_B$	Han_B	-0.039	-40.0	112,206	-0.041	-49.6	255,550
$Yoruba_B$	Ust'-Ishim	-0.029	-29.5	112,009	-0.030	-34.0	254,942
$Yoruba_B$	Kostenki14	-0.031	-29.4	105,169	-0.032	-37.0	240,458
$Yoruba_B$	MA1	-0.031	-27.3	81,023	-0.032	-34.5	186,013
$Yoruba_B$	Loschbour	-0.029	-30.9	111,229	-0.031	-38.9	252,674
$Yoruba_B$	LaBrana	-0.031	-29.2	107,988	-0.032	-37.0	246,902
$Yoruba_B$	Han_B	-0.031	-33.7	112,207	-0.032	-42.2	255,554
$Dinka_B$	Ust_Ishim	-0.024	-24.9	111,997	-0.026	-29.5	254,928
$Dinka_B$	Kostenki14	-0.026	-24.9	105,156	-0.027	-31.0	240,440
$Dinka_B$	MA1	-0.026	-22.5	81,018	-0.028	-29.9	186,004
$Dinka_B$	Loschbour	-0.025	-25.1	111,220	-0.027	-32.1	252,664
$Dinka_B$	LaBrana	-0.026	-24.1	107,976	-0.028	-30.6	246,888
$Dinka_B$	Han_B	-0.026	-27.1	112,207	-0.028	-34.0	255,554

Oase 1 has no evidence of affinity to other Europeans sampled to date

Extended Data 1 reports statistics of the form $D(\text{Non-African}_1, \text{Non-African}_2; \text{Oase 1, African})$ for all sites only; Extended Data Table 2 reports the same for transversions. Here, African refers to a pool of 6 genomes (2 Yoruba, 2 Dinka, and 2 Mbuti); East Asian to a pool of 4 genomes (2 Han and 2 Dai); and Native American to a pool of 3 genomes (2 Karitiana and 1 Mixe).

We observe that Oase 1 has no evidence of more allele sharing with ancient or present-day Europeans, than with non-Europeans, despite being from Europe. This can be seen by examining statistics of the form *D(European, non-European; Oase 1, African)*, where we represent non-European by any of Ust'-Ishim, East Asian, Native American, or MA1. We break this finding down into two classes.

Pre-agricultural Europeans

When the European sample is pre-agricultural (Kostenki 14 or Loschbour), the statistic is never significant after correcting for multiple hypothesis testing: $|Z| \leq 1.5$ for all sites (Fig. 1; Extended Data Table 1) and $|Z| \leq 2.3$ for transversions (Extended Data Table 2). This is not an issue of limited power, as when we perform the same analysis replacing Oase 1 with the nearly as old Kostenki 14 (using exactly the same number of SNPs), the scores are often highly significant. For example for all sites, *D(Loschbour, East Asian; X, African)* is $Z = -0.4$ when X = Oase 1, and $Z = 13.7$ when X = Kostenki14. Thus, Kostenki 14 has strong evidence of being on a lineage leading to later Europeans whereas Oase 1 has none.

Post-agricultural Europeans

When the European sample is post-agricultural (the ~7,000 years old early European farmer from Stuttgart [16] or a pool of four present-day Europeans), the statistics are all negative and sometimes significantly so: $-6.4 \leq Z \leq -2.3$ for all sites (Fig. 1; Extended Data Table 1) and $-4.8 \leq Z \leq -1.4$ for transversions (Extended Data Table 2). Thus, Oase 1 shares more alleles with non-Europeans than with post-agricultural Europeans, opposite to the expectation if there was genetic continuity between Oase 1 and later Europeans (or European contamination in Oase 1). A possible explanation for this observation is that post-agricultural Europeans have ancestry from Near Eastern migrants that brought agriculture to Europe, who in turn had ancestry from a population that diverged from pre-agricultural Europeans and non-Europeans before they separated from each other. Such ancestry, which has previously been suggested [6,16], would be expected to bias our statistic negative, as we observe. As would be predicted based on this explanation, a negative bias of a similar magnitude is seen when we replace Oase 1 with Ust'-Ishim in the statistic: $-5.9 \leq Z \leq -3.9$ for all sites (Extended Data Table 1) and $-5.2 \leq Z \leq -1.3$ for transversions (Extended Data Table 2).

It is possible that with more data from Oase 1, a signal of genetic continuity with later Europeans could be detected. However, it is interesting that to the limits of our resolution, the data are consistent with Oase 1 deriving from a lineage that went extinct in Europe, contributing little or nothing to subsequent populations (unlike Kostenki 14's population).

The *D*-statistic analyses also allow us make a more general statement about the relationship of Oase 1 to other modern human genomes analyzed to date.

Fig. 1, Extended Data Table 1, and Extended Data Table 2 show that Oase 1 shares alleles at an indistinguishable rate with diverse modern humans, including East Asians (a pool of 2 Han and 2 Dai), ancient Siberians (Ust'-Ishim), and other pre-agricultural Europeans (Kostenki 14 and Loschbour). The |Z|-scores in the top half of the table are ≤ 1.7 for Extended Data Table 1 and ≤ 2.3 for Extended Data Table 2, which is not significant after correcting for multiple hypothesis testing.

We conclude that a model that fits the data, to the limits of our resolution, is that the Oase 1 lineage separated from these other Eurasian lineages around the time of their divergence from each other.

Supplementary Information 4

On the order of six to nine percent Neanderthal ancestry in Oase 1

***D*-statistics**

We used *D*-statistics as implemented in the ADMIXTOOLS software[15] to test whether Oase 1 has a different proportion of Neanderthal ancestry than other ancient and present-day modern humans.

If *W, X, Y, Z* are 4 populations, and we randomly draw alleles in each population at each SNP *i* (which has non-reference allele frequencies w_i, x_i, y_i, and z_i in the four populations), then we are interested in two types of allele patterns:

$p_i(\mathrm{BABA}) = w_i(1-x_i)y_i(1-z_i) + (1-w_i)x_i(1-y_i)z_i$
the probability that *W* and *Y* match for one allele and *X* and *Z* for the alternate allele

$p_i(\mathrm{ABBA}) = w_i(1-x_i)(1-y_i)z_i + (1-w_i)x_iy_i(1-z_i)$
the probability that *W* and *Z* match for one allele and *X* and *Y* for the alternate allele

If we define

$E[n_{BABA}] = \sum_{i=1}^{n} p_i(BABA)$
the expected number of BABA sites over all SNPs in the dataset

$E[n_{ABBA}] = \sum_{i=1}^{n} p_i(ABBA)$
the expected number of ABBA sites over all SNPs in the dataset

We can then define

$$D(W,X;Y,Z) = \frac{E[n_{BABA}]-E[n_{ABBA}]}{E[n_{BABA}]+E[n_{ABBA}]}$$

If populations (*W, X*) descend from a common ancestral population since separation from (*Y, Z)*, the statistic should be consistent with 0. If there has been gene flow between either or both of the pairs (*W, Y*) or (*X, Z*) since separation from the others, the statistic will be positive. Similarly, if there has been gene flow between either or both of the population pairs (*W, Z*) or (*X, Y*) since separation from the others, the statistic will be negative. Thus, we can test the null hypothesis of (*W, X*) and (*Y, Z*) being clades by testing whether the statistic is consistent with zero.

We use a Weighted Block Jackknife[20] with a block size of 5 million base pairs (5 Mb) to compute standard errors, as implemented in ADMIXTOOLS.

For the analyses that follow, we pool data from Panels 1-3. After restricting to deaminated fragments, Oase 1 has 271,326 SNPs covered at least once, of which 118,938 are transversions. These numbers are not discrepant with the 242,122 SNPs reported in Extended Data Table 1, and the 106,005 SNPs reported in Extended Data Table 2, which correspond to the SNPs with coverage not just in Oase 1 but also in Kostenki 14, Ust'-Ishim, Loschbour and diverse present-day genomes.

Neanderthal ancestry in Oase 1

To determine whether Oase 1 has Neanderthal ancestry, we first computed the statistic *D*(*X, African; Altai Neanderthal, Chimp*) (Table S4.1). Oase 1 has evidence of more allele sharing with Neanderthals than with a pool of six sub-Saharan Africans (Z=7.7). The magnitude of the *D*-statistic is higher than that in other modern humans (D=0.0051 compared to D=0.0016-0.0031 for all others analyzed), suggesting the possibility that Oase 1 might have more Neanderthal ancestry than the others.

Table S4.1. ***D(X, African; Altai, Chimp)*** **restricted to transversions**

X	D	Z-score	Sites used
Oase1	0.0051	7.7	112,146
Ust'-Ishim	0.0025	8.1	1,035,603
Kostenki14	0.0031	9.7	913,271
MA1	0.0028	7.6	691,429
Loschbour	0.0022	7.3	1,030,375
Stuttgart	0.0022	7.9	1,025,311
Han_B	0.0021	6.9	1,037,648
Dai_B	0.0021	7.2	1,037,582
$French_B$	0.0016	6.1	1,037,637
$Sardinian_B$	0.0023	8.0	1,037,664
$Papuan_B$	0.0026	8.2	1,037,556

Table S4.2. ***D(Neanderthal$_1$, Neanderthal$_2$; Test, Outgroup).***

This analysis uses Panels 1-3 SNP data restricting to transversions.

			Chimp = Outgroup			Africa = Outgroup		
$Neand_1$	$Neand_2$	Test	D	Z	Sites	D	Z	Sites
Mezmaiskaya	Vindija	Oase 1	0.0005	0.5	11,854	0.0007	1.1	12,230
Mezmaiskaya	Vindija	Ust'-Ishim	0.0016	4.3	71,792	0.0007	3.1	73,776
Mezmaiskaya	Vindija	Kostenki 14	0.0013	3.2	67,039	0.0004	1.3	68,891
Mezmaiskaya	Vindija	MA1	0.0022	4.7	51,492	0.0012	3.8	52,912
Mezmaiskaya	Vindija	Loschbour	0.0018	4.9	71,339	0.0009	4.1	73,296
Mezmaiskaya	Vindija	LBK	0.0019	5.2	71,202	0.0010	4.5	73,156
Mezmaiskaya	Vindija	Han_B	0.0015	4.0	71,920	0.0006	2.8	73,908
Mezmaiskaya	Vindija	$French_B$	0.0017	4.6	71,918	0.0008	3.5	73,906
Mezmaiskaya	Altai	Oase 1	0.0012	1.4	17,297	0.0007	1.2	17,879
Mezmaiskaya	Altai	Ust'-Ishim	0.0028	9.1	113,097	0.0009	4.5	116,294
Mezmaiskaya	Altai	Kostenki 14	0.0026	7.6	104,617	0.0008	3.6	107,591
Mezmaiskaya	Altai	MA1	0.0032	8.1	80,054	0.0012	4.5	82,315
Mezmaiskaya	Altai	Loschbour	0.0031	10.0	112,348	0.0012	5.9	115,515
Mezmaiskaya	Altai	LBK	0.0032	10.0	112,090	0.0013	6.5	115,232
Mezmaiskaya	Altai	Han_B	0.0032	10.3	113,315	0.0013	6.6	116,521
Mezmaiskaya	Altai	$French_B$	0.0034	10.7	113,311	0.0015	7.4	116,519
Vindija	Altai	Oase 1	0.0009	2.7	75,540	0.0001	0.5	77,679
Vindija	Altai	Ust'-Ishim	0.0007	4.3	599,737	0.0000	0.3	614,897
Vindija	Altai	Kostenki 14	0.0008	4.9	541,699	0.0002	1.4	555,515
Vindija	Altai	MA1	0.0007	4.0	412,864	0.0001	0.7	423,304
Vindija	Altai	Loschbour	0.0007	4.7	596,711	0.0001	1.2	611,709
Vindija	Altai	LBK	0.0007	4.2	594,522	0.0000	0.4	609,482
Vindija	Altai	Han_B	0.0009	6.0	600,664	0.0003	2.8	615,882
Vindija	Altai	$French_B$	0.0009	5.8	600,649	0.0003	2.6	615,871

We also tested if archaic humans share more alleles with Oase 1 or with other non-Africans ("Test") using the statistic *D(Test, Oase 1; Archaic, Outgroup)*. Here, the Archaic individual is either the Altai Neanderthal or the Siberian Denisovan, and Outgroup is either chimpanzee or a pool of 6 sub-Saharan Africans. We restricted to transversions for this analysis in order to not be biased by the high rate of deamination in the archaic genomes. Extended Data Table 3 shows that Oase 1 shares significantly more derived alleles with the Neanderthal genome than with any other modern human individual tested, both when using the chimpanzee ($-3.6 \geq Z \geq -6.9$) and when using a Mbuti African ($-4.7 \geq Z \geq -8.2$) as outgroups. Using the Siberian Denisovan genome to represent the Archaic individual, the signal is weaker but present, as

expected for a scenario in which Oase 1 derives ancestry from Neanderthals, but not Denisovans (Denisovans are distantly related to Neanderthals).

We also tested if Oase 1 shares more derived alleles with a particular Neanderthal individual than with others. The affinity to different Neanderthals in Oase 1 is consistent with what has been observed in other modern humans studied to date, in the sense that the *D*-statistics are of consistent magnitude (Table S4.2). However, there is less data for Oase 1, and thus we may not have enough resolution to detect the differences in the Neanderthal population that contributed to Oase 1 compared to other samples, even if such differences exist.

Estimates of Neanderthal ancestry proportion using f_4-ratio statistics

To estimate the proportion of the Oase 1 genome that derives from Neanderthals, we use three different ratios of f_4-statistics[15] that exploit different parts of the historical relationships among the samples (Table 1 and Extended Data Table 4).

Statistic 1. The numerator is a quantity proportional to the correlation in the allele frequency difference between a test modern human and sub-Saharan African modern humans on the one hand, and Altai and Denisova on the other. We divide this by the same statistic substituting a test sample with the Mezmaiskaya Neanderthal[21]. This statistic assumes that the Neanderthal that introgressed into the ancestors of the tested modern human sample is a sister group to Mezmaiskaya; if this is wrong the estimate has the potential to be biased.

Statistic 2. This is computed as 1 minus an estimate of modern human ancestry. To obtain an estimate of modern human ancestry, we use a statistic whose numerator is proportional to the correlation in allele frequency difference between a test sample on the one hand and an archaic sample, and a sub-Saharan African and chimpanzee on the other. If the test sample is an archaic individual from the Neanderthal/Denisova clade, this statistic has an expectation of zero. We then divide by the same quantity replacing the test with Dinka, which is an approximate clade with non-African populations relative to other sub-Saharan Africans, so that the quantity is what is expected for a modern human with little or no Neanderthal ancestry[8]. An appealing feature of this statistic is that it works equally well if Altai or Denisova is used as the archaic (we used the Denisovan genome). The statistic also does not assume any relationship among the Neanderthals, contrasting with Statistic 1. This statistic is similar to Equation S8.5 of ref.[22].

Statistic 3. This statistic was introduced in ref.[22]. The numerator is proportional to the correlation in allele frequency difference between a test sample and a sub-Saharan African, and Denisovan with chimpanzee. We divide by the same quantity for a 100% archaic individual. The result has a higher standard error than Statistics 1 and 2, but has the appealing feature that it does not assume any relationships among the Neanderthals.

All three statistics indicate that Oase 1 has a higher proportion of Neanderthal ancestry than the other genomes tested. For all sites, the point estimates are 6.0% to 9.4%, and for Statistic 2 the lower bound of the ancestry estimate for Oase 1 excludes the upper bound for all other modern humans we analyzed (Table 1). For transversions only, the point estimates are 8.4% to 11.3%, and for both Statistic 1 and Statistic 2, the lower bound of the estimate for Oase 1 excludes the upper bound for all other modern humans we analyzed (Extended Data Table 4).

Supplementary Information 5

Oase 1 had a Neanderthal ancestor four to six generations back

SNPs indicative of Neanderthal ancestry
From the 1,749,385 SNPs targeted in the archaic probe set (Panel 4), we selected 954,849 SNPs where at least one Neanderthal allele differs from the majority of Yoruba (thus excluding SNPs differing only between the Denisovan genome and Yoruba).

We identified a total of 87,803 sites covered by at least one deaminated fragment in Oase at the first 5' and last two 3' bases in the UDG-treated libraries and the terminal three bases in the non-UDG-treated libraries (Extended Data Table 8). We then further restricted to sites that also had coverage in Ust'-Ishim, Kostenki 14, as well as in the Han, French and Dinka individuals from Panel B of Extended Data Table 9. This left 78,055 SNPs.

Oase 1 was always represented by a single allele (the majority call of the analyzed fragments) at each of these SNPs, which contrasted with the five other modern humans which were represented by high quality genomes with diploid genotype calls. To make the analyses comparable, we randomly sampled one of the two alleles for each of these five individuals. We then scored each SNP as 1 for carrying the Yoruba allele or 0 for the Neanderthal allele.

The sum of alleles matching Neanderthal for the 78,055 sites for each individual is shown in Extended Data Table 5. Oase 1 has a higher sum than the other individuals, consistent with having a higher proportion of Neanderthal ancestry (Supplementary Information section 4). The extent of this excess is highly variable across chromosomes (Extended Data Table 5).

The Dinka, a sub-Saharan African population, are thought to have little or no Neanderthal ancestry[21]. Thus, we hypothesized that we could interpret the 485 alleles matching Neanderthal in the Dinka individual as an estimate of the false-positive rate, that is, the fraction of sites in the panel of SNPs we are analyzing that are expected to carry the derived allele just by chance, without reflecting Neanderthal ancestry. We therefore subtracted the number of derived alleles in Dinka from that in the other individuals, and hypothesized that the residual rate of derived alleles in the other genomes is proportional to Neanderthal ancestry. Assuming that the French individual has 2.0% Neanderthal ancestry, we infer 7.3% Neanderthal ancestry in Oase 1, 3.1% in Ust'-Ishim, and 2.6% in Han. These numbers are similar to those in Supplementary Information section 4, Table 1, and Extended Data Table 4, and continue to support the finding of more Neanderthal ancestry in Oase 1.

Large stretches of Neanderthal ancestry in Oase 1
Fig. 2 shows the physical distribution across the autosomes (chromosomes 1-22) of alleles where a randomly sampled allele from a modern human (Oase 1, Ust'-Ishim, Kostenki 14, Han, French and Dinka) matches Neanderthal rather than the majority of Yoruba.

It is visually evident that there are large stretches of the autosomes where there are high rates of alleles matching Neanderthal. For example, we observe segments of high rates of Neanderthal matching on chromosome 5 and chromosome 12 that are at least 50 million base pairs (Mb) in size. These stretches are far larger even than the largest segments reported previously in Ust'-Ishim which are up to 6 Mb[6].

To identify the endpoints of likely large stretches of Neanderthal ancestry in Oase 1, we wrote an algorithm for calling chunks. The core machinery of this algorithm is a Hidden Markov Model (HMM), but no care is taken to make the parameters meaningful from a probabilistic point of view, so the posterior decoding probabilities that emerge should not be literally interpreted as estimates of introgression probability. The algorithm works as follows:

(1) Bin each of the chromosomes into 2 million base pair non-overlapping chunks.

(2) Restrict analysis to bins with at least 20 sites with data. This leaves 1,314 bins.

(3) Within each bin in each individual, report a single number, the fraction f of SNPs in the bin where the allele representing that individual is derived.

(4) Run an HMM using the same engine as described in Supplementary Information section 13 of [21], where:

t = minimum f for which we view the chunk as giving evidence of introgression
p_{nea} = probability that the chunk is Neanderthal introgressed conditional on $f \geq t$
p_{mod} = probability that the chunk is Neanderthal introgressed conditional on $f < t$
s = switch rate between Neanderthal and modern human chunks per base pair
q = prior probability that the chunk is Neanderthal introgressed

We ran this procedure over a grid of values of t, p_n, p_m and s, and manually chose a parameter combination that resulted in a high fraction of the genome being called with confidence as either modern human or Neanderthal. The parameters used were:

$t = 0.1$
$p_{nea} = 0.8$
$p_{mod} = 0.05$
s = 20,000,000 base pairs
$q = 0.1$

This produced a posterior decoding in which 90.9% of the genome had a value of <0.1 (little evidence of a large chunk of Neanderthal ancestry), and 7.5% >0.8.

The Oase 1 "smoothed" track in Figure 1 shows yellow coloring at each position where the posterior decoding has a value of >0.5. The seven intervals are given in Table S5.1. We make no claim that these are the only segments of very recent Neanderthal introgression in Oase 1. We only claim that they are some of the most easily recognizable.

Table S5.1. Coordinates of seven segments of very recent Neanderthal introgression

Chromosome	Start	Stop	Physical span (Mb)	Genetic span (cM)
4	90.094	115.953	25.859	22.329
5	85.046	147.953	62.908	53.507
6	6.005	17.957	11.952	20.447
6	72.044	88	15.956	9.229
9	0.257	27.978	27.721	52.801
9	88.025	113.983	25.958	30.058
12	26.037	91.946	65.909	57.703

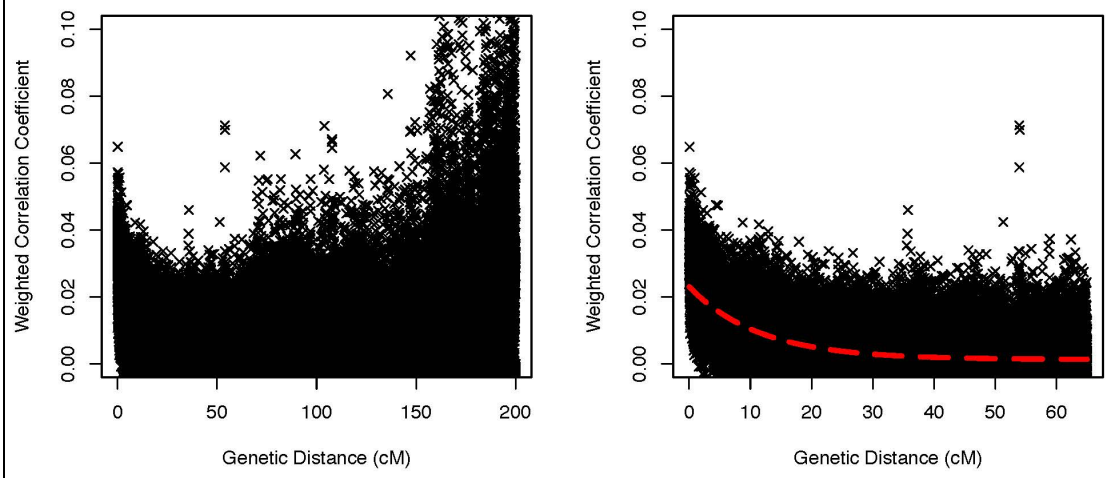
Dating the most recent Neanderthal admixture into the ancestors of Oase 1

Admixture between populations induces correlation in ancestry across the genome of an admixed individual, and the extent of this correlation is informative about the time since mixture [23-25]. As in[6] we use this signal to estimate the date of Neanderthal introgression.

Method A – Fitting an exponential decay

We use the Oxford combined genetic map[26] and calculate the average covariance over all pairs of SNPs in 0.001 cM bins that carry Neanderthal alleles as defined above. The exponential decay gets lost in noise around 65 cM (Figure S5.1) and thus we fit to this point.

Figure S5.1. Pairwise covariance for SNPs that match the ascertainment scheme in which Neanderthal carries different alleles from Africans. The decay gets lost in noise beyond around 65 cM (left panel) so we only fit an exponential function up to this point (right panel).



Assuming a single pulse of admixture, we fit an exponential function ($y = Ae^{-(n-1)d} + c$), where n = number of generations since Oase 1 had a Neanderthal ancestor and d = genetic distance (in Morgans). Figure S5.1 shows the covariance curves. We compute standard errors using a Weighted Block Jackknife[20], removing one chromosome in each run and studying the variability in the estimated dates of mixture. Using this method, we estimate that Oase 1 had Neanderthal ancestors 8.1 ± 5.5 generations back in his family tree.

This estimated date of admixture is an average of both the most recent Neanderthal admixture into the ancestors of Oase 1, and older admixture that was perhaps shared with other non-Africans. We did not succeed in fitting the data as a mixture of two exponential distributions. We therefore turned to alternative methods that focused on dating the most recent mixture.

Method B – Fitting the distribution of large chunks

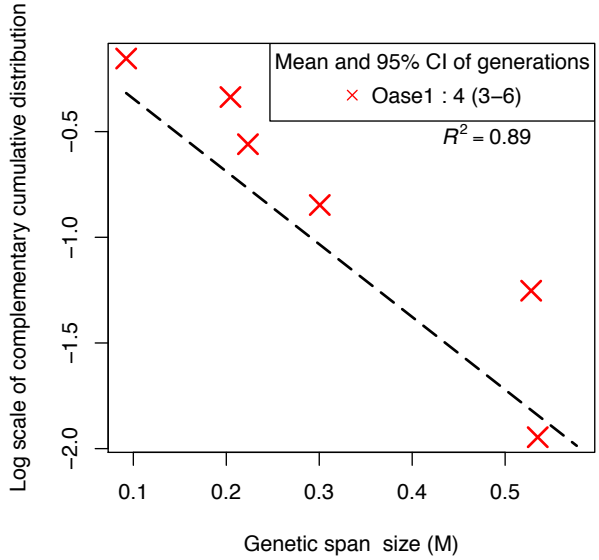
Because the chunks of Neanderthal introgression are so large that we can reasonably infer the positions of the largest chunks, and because we are interested specifically in the admixture that gave rise to the largest chunks, we reasoned that it might be valid to simply determine the sizes of the large chunks, and then to fit a distribution to them. For this analysis, we use the seven chunks identified by the HMM, whose positions are listed in Table S5.1.

We assume that the chunks resulting from the most recent admixture event have a distribution $y = Ae^{-(n-1)d}$, where n = number of generations since Oase 1 had a Neanderthal ancestor and d = genetic distance (in Morgans). The fact that the exponent has the quantity

(*n-1*) in it reflects the fact that Neanderthal and modern human chromosomes are only expected to begin to be observed in recombinant form in the second generation after admixture (in the first generation after admixture, each individual has one entirely Neanderthal and one entirely modern human chromosome at each locus).

Taking the natural logarithm of this distribution, we get $ln(y) = -(n-1)d + ln(A)$. Figure S5.2 shows the cumulative distribution of the number of chunks, plotted on a log-scale to allow a linear fit. The slope of this plus 1 translates to a date of admixture. The resulting fit is good (R^2=0.89), and has a slope corresponding to an admixture date of 4 generations ago (95% confidence interval obtained by linear regression of 3-6 generations ago).

Figure S5.2 Fraction of the Oase 1 genome comprised of the seven largest chunks. The slope of the curve is expected to be (n-1), and can be used to estimate the number of generations since Oase 1 had his most recent Neanderthal ancestor.



Method C - Probability of the observed spans of the largest Neanderthal chunks

We examined the expected distribution of the largest chunks of introgression, including the 1 1argest, 2 largest, 3 largest, and 4 largest, for different time depths of introgression. We reasoned that the genetic spans of the largest chunks might be robust statistics because:

- We likely have good power to recognize the largest chunks, but may have imperfect power to recognize the smaller chunks. Thus, fitting a model to the largest chunks may be more robust than fitting a model including some smaller chunks. As chunk sizes become smaller, it becomes increasingly possible that we have missed chunks of similar size that are real and hence our measurement of the distribution is less likely to be accurate.

- We are interested here in studying the most recent Neanderthal introgression. When we include shorter chunks in the analysis, we expect to see a size range where older Neanderthal introgression events may be contributing to the observed patterns, and thus our estimate of the date may be higher than that of the most recent mixture.

We carried out a series of simulations that fragmented the genome generation by generation assuming the empirically measured sex-averaged genetic map (using the map from[27]).

Table S5.2 reports the results of this simulation study, showing the number of generations since mixture that produces a 95% confidence interval of average chunk size for the top *k* chunks that includes our empirical observation. We observe that for just the top chunk (*k*=1), Oase 1 is consistent with having Neanderthal admixture 4-8 generations in the past. For averages of the top 1-2 (*k*=2), 1-3 (*k*=3), and 1-4 (*k*=4) chunks, Oase 1 is consistent with having Neanderthal admixture 4-6 generations ago. We conclude that the most recent admixture likely occurred 4-6 generations ago.

Table S5.2. Average size of the *k* largest chunks for different numbers of generations since mixture. We carried out 10,000 simulations for each parameter combination (number of generations since mixture, and number *k* of the largest chunks being averaged), and give the 95% credible interval. Gray boxes include the empirical value.

Generation	k=1	k=2	k=3	k=4
1	270-271	263-264	248-249	237-238
2	128-271	122-231	115-210	111-196
3	72-206	67-172	63-152	59-138
4	41-159	37-127	34-109	31-97
5	22-122	19-95	16-80	15-70
6	9-95	8-71	7-59	7-51
7	2-75	3-54	4-44	5-39
8	1-60	2-43	3-36	4-32
9	0-49	1-35	2-30	3-27
10	0-42	1-31	2-26	3-24
Empirical	58	56	55	49

We can compare the estimated date of Neanderthal mixture obtained from the chunk size distribution to that needed to produce the observed genome-wide proportion of Neanderthal ancestry if all the Neanderthal admixture occurred recently. The average proportions of Neanderthal ancestry expected to be contributed by ancestors at different time depths of admixture are given in Table S5.3. Assuming that all the extra Neanderthal ancestry in Oase 1 was contributed by a single recent ancestor, it would be expected from Table S5.3 that this ancestor lived 3-5 generations back. This overlaps with the estimates from chunk span estimates: 3-6 generations back by Method B and 4-6 generations back by Method C.

Table S5.3. Average proportion of Neanderthal ancestry expected to be contributed by a Neanderthal ancestor different numbers of generation ago

Generations back	Ancestor	Proportion Neanderthal
1	Parent	50%
2	Grandparent	25%
3	Great-grandparent	12.5%
4	Great-great-grandparent	6.25%
5	Great-great-great-grandparent	3.13%
6	Great-great-great-great-grandparent	1.56%

We conclude that it is likely that Neanderthal admixture happened 4-6 generations ago. In other words, Oase 1 possibly had a Neanderthal great-great-grandparent and certainly had a Neanderthal great-great-great-great-grandparent.

Supplementary References

1 Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2223-2227 (2013).

2 Renaud, G., Kircher, M., Stenzel, U. & Kelso, J. freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics* **29**, 1208-1209 (2013).

3 Meyer, M. *et al.* A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* **505**, 403-406 (2014).

4 Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Current biology : CB* **23**, 553-559 (2013).

5 Shapiro, B. *et al.* A Bayesian phylogenetic method to estimate unknown sequence ages. *Molecular biology and evolution* **28**, 879-887 (2011).

6 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449 (2014).

7 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).

8 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226 (2012).

9 Green, R. E. *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416-426 (2008).

10 Kloss-Brandstatter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human mutation* **32**, 25-32 (2011).

11 Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708-713 (2000).

12 Briggs, A. W. *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318-321 (2009).

13 Posada, D. & Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-818 (1998).

14 Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214 (2007).

15 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).

16 Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409-413 (2014).

17 Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* (2015).

18 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722 (2010).

19 Trinkaus, E. *et al.* An early modern human from the Pestera cu Oase, Romania. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11231-11236 (2003).

20 Kunsch, H. R. The Jackknife and the Bootstrap for General Stationary Observations. *Annals of Statistics* **17**, 1217-1241 (1989).

21 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49 (2014).

22 Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010).
23 Loh, P. R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233-1254 (2013).
24 Moorjani P, P. N., Hirschhorn JN, Keinan A, Hao L, et al. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. . *PLoS Genet* **7**, e1001373 (2011).
25 Hellenthal, G. *et al.* A Genetic Atlas of Human Admixture History. *Science* **343**, 747-751 (2014).
26 International HapMap, C. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
27 Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature genetics* **31**, 241-247 (2002).