nature COMMUNICATIONS

ARTICLE

Received 4 Aug 2015 | Accepted 9 Dec 2015 | Published 19 Jan 2016

DOI: 10.1038/ncomms10408 OPEN

Iron Age and Anglo-Saxon genomes from East England reveal British migration history

Stephan Schiffels[1,†], Wolfgang Haak[2,†], Pirita Paajanen[1,†], Bastien Llamas[2], Elizabeth Popescu[3], Louise Loe[4], Rachel Clarke[3], Alice Lyons[3], Richard Mortimer[3], Duncan Sayer[5], Chris Tyler-Smith[1], Alan Cooper[2] & Richard Durbin[1]

British population history has been shaped by a series of immigrations, including the early Anglo-Saxon migrations after 400 CE. It remains an open question how these events affected the genetic composition of the current British population. Here, we present whole-genome sequences from 10 individuals excavated close to Cambridge in the East of England, ranging from the late Iron Age to the middle Anglo-Saxon period. By analysing shared rare variants with hundreds of modern samples from Britain and Europe, we estimate that on average the contemporary East English population derives 38% of its ancestry from Anglo-Saxon migrations. We gain further insight with a new method, rarecoal, which infers population history and identifies fine-scale genetic ancestry from rare variants. Using rarecoal we find that the Anglo-Saxon samples are closely related to modern Dutch and Danish populations, while the Iron Age samples share ancestors with multiple Northern European populations including Britain.

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [2] Australian Centre for Ancient DNA, School of Biological Sciences and The Environment Institute, University of Adelaide, Adelaide, South Australia 5005, Australia. [3] Oxford Archaeology East, 15 Trafalgar Way, Bar Hill, Cambridge CB23 8SQ, UK. [4] Oxford Archaeology South, Janus House, Osney Mead, Oxford OX2 0ES, UK. [5] School of Forensic and Applied Sciences, University of Central Lancashire, Preston PR1 2HE, UK. † Present addresses: Department for Archaeogenetics, Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, 07745 Jena, Germany (S.S. or W.H.); The Genome Analysis Centre, Norwich Research Park, Norwich NR4 7UH, UK (P.P.). Correspondence and requests for materials should be addressed to S.S. (email: schiffels@shh.mpg.de) or to R.D. (email: rd@sanger.ac.uk).

Within the last 2,000 years alone, the British Isles have received multiple well-documented immigrations. These include military invasions and settlement by the Romans in the first century CE, peoples from the North Sea coast of Europe collectively known as the Anglo-Saxons between ca. 400 and 650 CE (Fig. 1a), Scandinavians during the late Saxon 'Viking period' 800–1,000 CE and the Normans in 1,066 CE (ref. 1). These events, along with prior and subsequent population movements, have led to a complex ancestry of the current British population. Although there is only a slight genetic cline from north to south at a coarse level[2,3], recent analyses have revealed considerable fine-scale genetic structure in the Northern and Western parts of Great Britain, alongside striking homogeneity in Southern and Eastern England[4] in the regions where archaeologists identify early Anglo-Saxon artifacts, cemeteries and communities. A variety of estimates of the fraction of Anglo-Saxon genetic ancestry in England have been given[5–8], with the recent fine structure analysis suggesting most likely 10–40% (ref. 4).

However, even large-scale analyses of present-day data provide only weak evidence of the Anglo-Saxon migration impact, mainly for two reasons. First, estimating the impact of historical migrations from present-day genetic data alone is challenging, because both the state of the indigenous population before the migration as well as the genetic make up of the immigrants are unknown and have to be estimated simultaneously from present-day data. Second, if the source population is genetically close to the indigenous population, migrations are hard to quantify due to the challenge in detecting small genetic differences. This is particularly true for the case of the Anglo-Saxon migrations in Britain, given the close genetic relationships across Europe[9,10].

Here we address both of these challenges using ancient DNA and new methodology. We present whole-genome sequences of 10 ancient samples from archaeological excavations in East England, which date to the late Iron Age and to the early and middle Anglo-Saxon periods and hence let us directly observe and quantify the genetic impact of the Anglo-Saxon migrations in England. Furthermore, we develop new methodology based on rare genetic variation in hundreds of modern samples to detect subtle genetic differentiation between immigrant and indigenous ancestry. We estimate that the modern-day East English population derives on average 38% of its ancestry from Anglo-Saxon migrants. We give evidence for mixing of migrants and natives in the early Anglo-Saxon period, and we show that the Anglo-Saxon migrants studied here have close ancestry to modern-day Dutch and Danish populations.

Results

Samples and sequencing. We generated genome sequences for 10 samples that were collected from three sites in East England close to Cambridge: Hinxton (five samples, Supplementary Fig. 1), Oakington (four samples, Supplementary Fig. 2) and Linton (1 sample), which were selected from a total of 23 screened samples based on DNA preservation (Fig. 1b, Table 1, Supplementary Table 1, Supplementary Note 1). All sequenced samples were radiocarbon dated (Supplementary Table 2), and fall into three time periods: the Linton sample and two Hinxton samples are from the late Iron Age (~ 100 BCE), the four samples from Oakington from the early Anglo-Saxon period (fifth to sixth century), and three Hinxton samples from the middle Anglo-Saxon period (seventh to ninth century; Fig. 1c). The two Iron Age samples from Hinxton are male, all other samples are female, based on Y chromosome coverage and consistent with the archaeology. All samples were sequenced to genome-wide coverage from 1x to 12x (Table 1). All have contamination rates below 2%, as estimated both from mitochondrial DNA and from nuclear DNA (Supplementary Table 3, Supplementary Note 2).

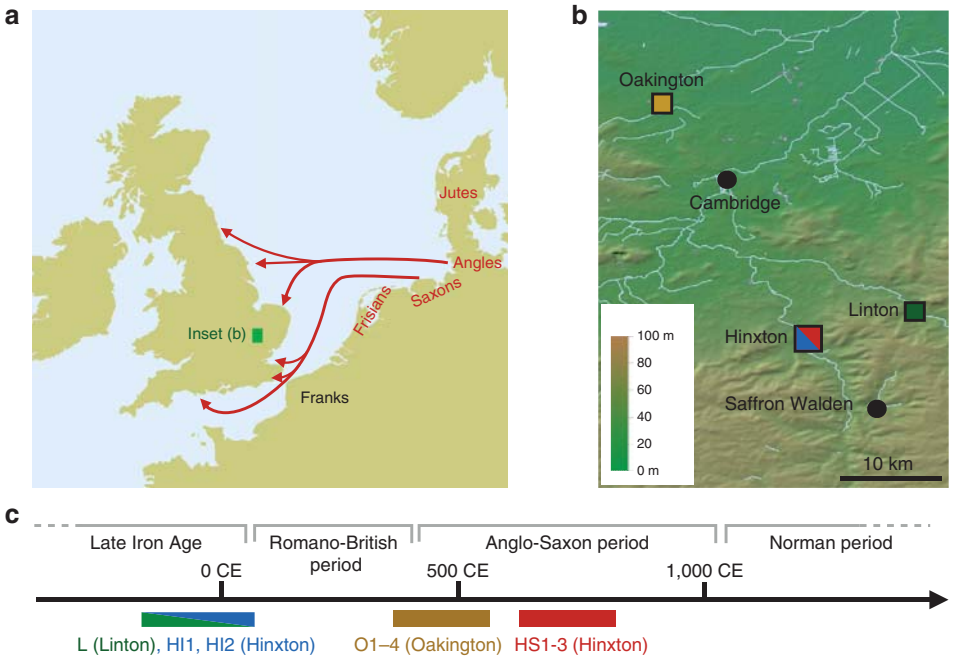


Figure 1 | Geographic and temporal context of the samples used in this study. (**a**) Anglo-Saxon migration routes of people from the continental coast, as reconstructed from historical and archaeological sources. (**b**) The ancient samples used in this study were excavated at three archaeological sites in East England: Hinxton, Oakington and Linton. The towns Cambridge and Saffron Walden are also shown (black circles). Background green/brown shades indicate altitude. The colours of the four sample match the ones in **c** and Fig. 2. (**c**) The 10 ancient samples belong to three age groups. The sample from Linton and two samples from Hinxton are from the late Iron Age, the four Oakington samples from the early Anglo-Saxon period and three Hinxton samples are from the middle Anglo-Saxon period.

Table 1 | A summary of all sequenced samples in this study.

Name	Origin	Sex	C14 Date (calibrated)	Endogenous (%)	Unique (%)	MT and Y haplogroup	Mean autosomal coverage
L	Linton	Female	360–50 BCE	72	54	H1e	1.4
HI1	Hinxton	Male	160 BCE–26 CE	16	63	K1a1b1b, R1b1a2a1a2c	1.3
HI2	Hinxton	Male	170 BCE–80 CE	83	65	H1ag1, R1b1a2a1a2c1	11.8
O1	Oakington	Female	420–570 CE	81	50	U5a2a1	3.8
O2	Oakington	Female	385–535 CE	92	68	H1g1	2.7
O3	Oakington	Female	395–540 CE	95	64	T2a1a	8.2
O4	Oakington	Female	400–545 CE	67	77	H1at1	6.3
HS1	Hinxton	Female	666–770 CE	36	91	H2a2b1	4.4
HS2	Hinxton	Female	631–776 CE	42	74	K1a4a1a2b	3.8
HS3	Hinxton	Female	690–881 CE	16	71	H2a2a1	0.9

The '% endogenous' values give the percentage of sequenced DNA that map to the human reference genome. The '% unique' values give the fraction of mapped reads that are left when excluding duplicates. The 'mean autosomal coverage' is the number of reads covering a base, averaged across chromosome 20. C14 Dates are calibrated, with 95% confidence intervals given.

Mitochondrial and Y chromosome haplogroups of all samples are among the most common haplogroups in present-day North-Western Europe (Table 1)[11,12] and in this case not informative for distinguishing immigrant versus indigenous ancestry.

We generated a principal component plot of the 10 ancient samples together with relevant European populations selected from published data[13,14] (Supplementary Fig. 3). The ancient samples fall within the range of modern English and Scottish samples, with the Iron Age samples from Hinxton and Linton falling closer to modern English and French samples, whereas most Anglo-Saxon era samples are closer to modern Scottish and Norwegian samples. Overall, though, population genetic differences between these samples at common alleles are small.

Estimating the Anglo-Saxon component in modern Britain. While principal component analysis can reveal relatively old population structure, such as generated from long-term isolation-by-distance models[15], whole-genome sequences let us study rare variants to gain insight into more recent population structure. We identified rare variants with allele frequency up to 1% in a reference panel of 433 European individuals from modern Finland, Spain, Italy, Netherlands and Denmark, for which genome-wide sequence data are available[16–18]. We determined for each ancient sample the number of rare variants shared with each reference population (Supplementary Note 3). There are striking differences in the sharing patterns of the samples, illustrated by the ratio of the number of rare alleles shared with Dutch individuals to the number shared with Spanish individuals (Fig. 2a). The middle Anglo-Saxon samples from Hinxton (HS1, HS2 and HS3) share relatively more rare variants with modern Dutch than the Iron Age samples from Hinxton (HI1 and HI2) and Linton (L). The early Anglo-Saxon samples from Oakington are more diverse with O1 and O2 being closer to the middle Anglo-Saxon samples, O4 exhibiting the same pattern as the Iron Age samples, and O3 showing an intermediate level of allele sharing, suggesting mixed ancestry. The differences between the samples are highest in low-frequency alleles and decrease with increasing allele frequency. This is consistent with mutations of lower frequency on average being younger, reflecting more recent distinct ancestry, compared with higher frequency mutations reflecting older shared ancestry.

We also examined using the same method 30 modern samples from the UK10K project[19], 10 each with birthplaces in East England, Wales and Scotland. Overall, these samples are closer to the Iron Age samples than to the Anglo-Saxon era samples (Fig. 2a). There is a small but significant difference between the mean values in the three modern British sample groups, with East English samples sharing slightly more alleles with the Dutch, and Scottish samples looking more like the Iron Age samples.

To quantify the ancestry fractions, we fit the modern British samples with a mixture model of ancient components, by placing all the samples on a linear axis of relative Dutch allele sharing that integrates data from allele counts 1–5 (Fig. 2b, Supplementary Note 3). By this measure the East England samples are consistent with 38% Anglo-Saxon ancestry on average, with a large spread from 25 to 50%, and the Welsh and Scottish samples are consistent with 30% Anglo-Saxon ancestry on average, again with a large spread (Supplementary Table 4). These numbers are lower on average if we exclude the low-coverage individual HS3 from the Anglo-Saxon group (35% for East English samples). A similar result is obtained when we analyse modern British samples from the 1,000 Genomes Project, which exhibit a strong substructure (Supplementary Note 4, Supplementary Fig. 4). We find that samples from Kent show a similar Anglo-Saxon component of 37% when compared against Finnish and Spanish outgroups, with a lower value for samples from Cornwall (Supplementary Fig. 5a, Supplementary Table 4).

An alternative and potentially more direct approach to estimate these fractions is to measure rare allele sharing directly between the modern British and the ancient samples. While being much noisier than the analysis using Dutch and Spanish outgroups, this yields consistent results (Supplementary Fig. 5b, Supplementary Note 3). In summary, this analysis suggests that on average 25–40% of the ancestry of modern Britons was contributed by Anglo-Saxon immigrants, with the higher number in East England closer to the immigrant source. The difference between groups within Britain is surprisingly small compared with the large differences seen in the ancient samples. This is true for both the UK10K samples and for the British samples from the 1,000 Genomes project, although we note that the UK10K sample locations may not fully reflect historical geographical population structure because of recent population mixing.

One caveat of our analysis is that we are using the three Iron Age samples from Cambridgeshire as proxies for the indigenous British population, which no doubt was structured, though it seems reasonable to take these as representatives at least for Eastern England. Furthermore, any continental genetic contribution from the Romano-British period would be factored into the assigned Anglo-Saxon component, as would a late Anglo-Saxon Scandinavian or Norman contribution. However these effects would only be strong if the contribution was large and heavily biased on the Dutch–Spanish axis.

Building a population history model from rare variants. To get further insight into the history underlying these sharing patterns, we developed a sensitive new method, rarecoal, which fits a demographic model to the joint distribution of rare alleles in a large number of samples (Supplementary Notes 5 and 6). Our

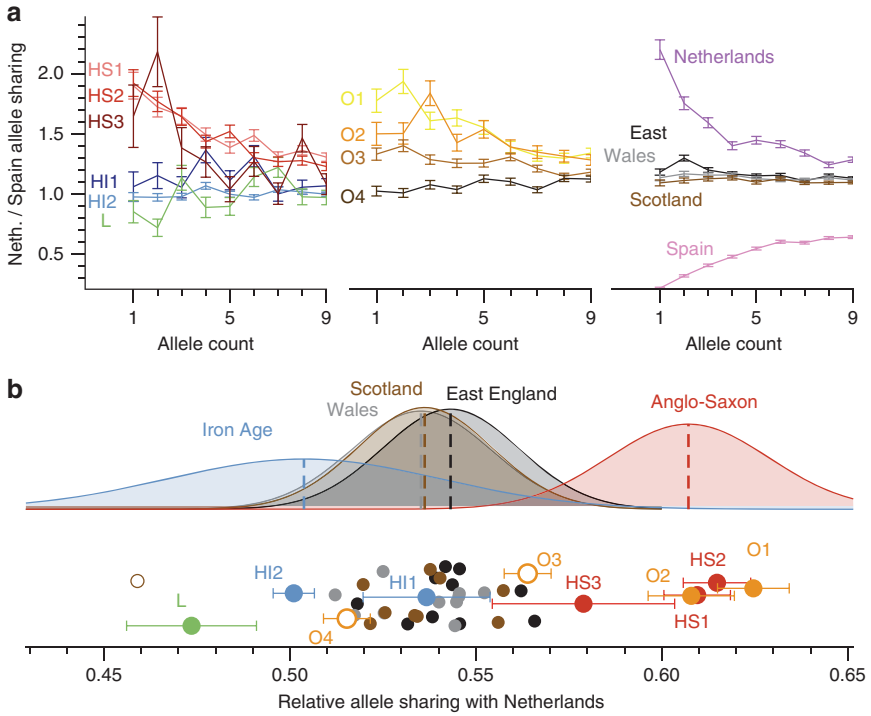


Figure 2 | Relative rare allele sharing between ancient and modern samples. (**a**) The ratio of the numbers of rare alleles shared with modern Dutch and Spanish samples as a function of the allele count in the set of modern samples. Ancient sample codes (left-hand and middle sections) are defined in Table 1. Results from present-day British individuals (right hand panel) are averaged over 10 individuals from each subpopulation. Results from a Dutch and a Spanish individual are shown for comparison. Error bars are calculated from raw count statistics and using s.e. propagation (Methods section). (**b**) The relative fraction of rare alleles shared with modern Dutch compared with Spanish alleles, integrated up to allele count five in the modern samples. Iron Age and Anglo-Saxon samples mark the two extremes on this projection, while modern samples are spread between them, indicating mixed levels of Anglo-Saxon ancestry, which is on average higher in East England than in Wales and Scotland, with a large overlap. Two Early Anglo-Saxon samples from Oakington have been excluded from computing the average, indicated by empty circles, because they show evidence for being admixed (O3) or of non-immigrant ancestry (O4). One modern sample from Scotland is also excluded, indidated as empty circle because it is a clear outlier with respect to all other Scottish samples. Samples are shown with a random vertical offset for better clarity. Error bars (Methods section) for the modern samples are omitted here, but of the same order of magnitude as for the ancient samples. Data for this figure is available as Supplementary Data 1.

strategy is to build a model in the form of a population phylogeny of the relationship between modern European populations, into which we can place the ancient samples. We recognize that a model without admixture and post-split gene flow is inadequate as a complete description of European population history. However, this is a natural simplified model, and the focus in this study is on understanding the genetic relationships of immigrants and indigenous populations in England, for which this population phylogeny model provides a reasonable scaffold.

The key idea is to model explicitly the uncertainty in the past of the distribution of derived alleles, but approximate the corresponding distribution for non-derived alleles by its expectation (Fig. 3a). Because rarecoal explicitly models rare mutations, it estimates separations in mutation clock time rather than genetic drift time, in contrast to methods based on allele frequency changes in common variants[20]. We first tested rarecoal on simulated data and found that it was able to reconstruct split times and branch population sizes with good accuracy (Fig. 3b), matching allele sharing almost exactly (Supplementary Fig. 6). We also tested its robustness with a smaller sample size in only one population (as in the Danish samples studied here), and under admixture (Supplementary Note 5, Supplementary Fig. 7).

We next applied rarecoal to 524 samples from six populations in Europe (Fig. 3c,d) to estimate a European demographic tree into which we could place the ancient samples. Because the British samples in the 1,000 Genomes Project fall into three distinct clusters, reflecting three sample locations (from Kent, Cornwall and the Orkney Islands, as part of the Peoples of the British Isles project[4,21], Supplementary Note 4)[16], we fitted different trees to these different groups (Supplementary Fig. 8). The common feature in all three trees is a first split between Southern and Northern Europe with a median time ~7,000 years ago, followed by three more separations close in time ~5,000 years ago between Netherlands, Denmark, Finland and Britain. Interestingly, when using the British samples from Cornwall, we obtained a tree where Cornwall forms an outgroup to the Dutch, Danish and Finnish population (Fig. 3c). In contrast, when we use Kent, it forms a clade with the Dutch population (Fig. 3d), consistent with higher Anglo-Saxon ancestry in the South of England than in Cornwall. When we use the Orkney population as the British branch, we find a similar tree topology as the one for Cornwall. These results show that both Cornwall and Orkney are more distantly related to continental Europe than Kent is. The tip branch effective population size is lowest in Finland (~12,000), consistent with previous observations[22,23], and highest in Kent (~191,000) and in the Netherlands (~184,000). For the European data, the allele sharing fit is worse than for the simulated data (Supplementary Fig. 9), presumably due to simplifying model assumptions of a constant population size in each branch and the absence of migration.

The relatively recent estimate for the split time between Italy and Spain, ~2,600 years ago, may be a consequence of migration following an earlier separation; the population size of the Italian-Spanish ancestral population was estimated to be extremely large

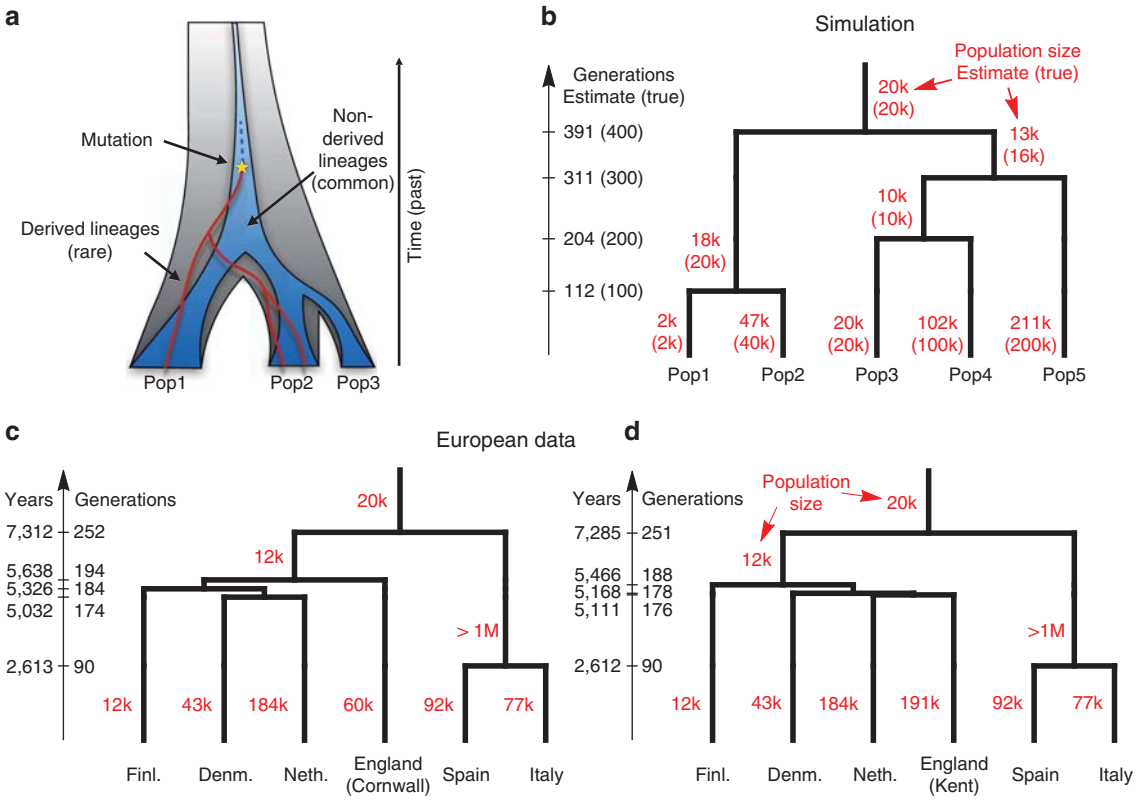


Figure 3 | Modelling European history with rarecoal. (**a**) Rarecoal tracks the probabilities for the lineages of rare alleles (red) within a coalescent framework back in time, and approximates the distribution of non-derived alleles (dark blue) by its average. (**b**) By optimizing the likelihood of the data under the model, we can estimate population sizes and split times. Tested with simulated data, the estimates closely match the true values (in parentheses). (**c**) Applied to hundreds of European individuals, rarecoal estimates split times as indicated on the time axis and population sizes for each branch. (**d**) Same as **c**, but using samples from Kent instead of Cornwall as a proxy for the British population. The different tree topology between **c** and **d** reflects different population histories in Cornwall compared with Kent in the South of England.

and an upper bound could not be determined, which could be an artifact of ancestral substructure or admixture. Another explanation would be a common source of admixture into both the Spanish and the Italian population, resulting in relatively recent common ancestry. We show in Supplementary Fig. 7 how admixture can modify rarecoal estimates of effective population size estimates and split times.

Modelling ancestry of ancient genomes using rarecoal. In addition to reconstructing the broader European relationship from a large sample set, rarecoal can be used to evaluate the relationship of a single ancient sample with the European tree. To do this, we assume a model in which the ancestral population of the single sample merges with the European tree at a particular branch at a particular time before the date of origin of the sample. We can then use rarecoal to evaluate the likelihood of the joint allele sharing data between the ancient sample and the modern populations under each model, specified by the branch and merge time in the tree (Fig. 4, Supplementary Note 5). There was a marked difference between the Iron Age and the Anglo-Saxon era samples: the Anglo-Saxon era samples mostly merged onto the Dutch and Danish branches, whereas the Iron Age samples preferentially merged at the base of the ancestral branch for all modern Northern European samples. The exception is that the early Anglo-Saxon O4 shows the same signal as the Iron Age samples, consistent with the rare allele sharing analysis (Fig. 2). For sample O3, which appeared to be of mixed ancestry in the allele sharing analysis, we find highest likelihood for merging with the Danish branch. However, in this sample there is also a notably higher likelihood to merge onto the same Northern European ancestral branch point as seen for the Iron Age samples. This is consistent with O3 being of recently mixed indigenous and Anglo-Saxon origin, although we can not rule out more complex scenarios involving prior mixed ancestry of this individual during the Romano-British period. There is some differentiation amongst the Anglo-Saxon era samples with samples O1, O2, HS1 and HS3 having highest likelihood of merging onto the Dutch branch while O3 and HS2 have highest likelihoods of merging onto the Danish branch, although in some cases the difference in likelihood between these two possibilities is small. The signals from HS3, HI1 and L are more spread due to low coverage, but consistent with the other results.

The mapping of the ancient samples onto the tree is similar for the tree using Kent as British population (Supplementary Fig. 10) and for the tree using Cornwall as the British proxy (Fig. 4). In particular, the Iron Age samples map onto the ancestral branch of Northern European populations irrespective of using Kent or Cornwall as British proxy. This suggests that none of the present-day populations in our data set, including the population from Cornwall, are as closely related to the Iron Age samples as Denmark and the Netherlands are to the Anglo-Saxon samples.

We validated our approach of mapping individual samples into a tree by placing modern samples onto the same tree as in Fig. 4. We find all samples from populations used in building the tree placed on the tip of their respective branch as expected (Supplementary Fig. 11). When mapping samples from groups not present in the tree, as is the case for samples from Kent and

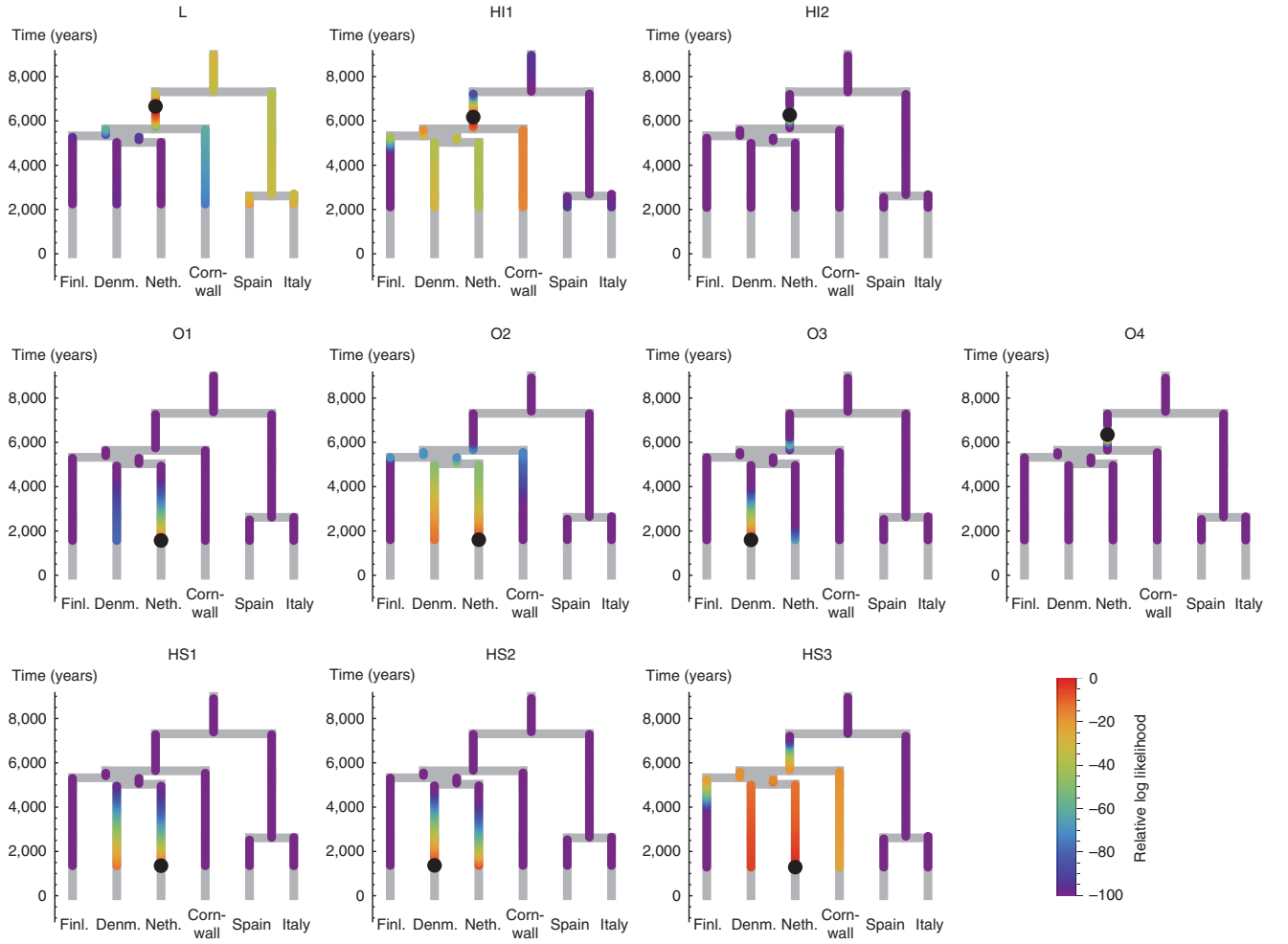


Figure 4 | Placing ancient samples into the European tree. Given the European tree with Cornwall as British population branch, we map ancient samples onto this tree. We colour each point in the tree according to the likelihood that the ancestral branch of the ancient sample merges at that point. The maximum likelihood merge point is marked by a black circle. The analysis shows that Iron Age samples L, HI1 and HI2 have highest likelihood to merge onto the ancestral branch of all Northern European populations analysed, whereas the Anglo-Saxon samples merge into the Dutch and Danish branches, respectively. The low coverage samples L, HI1 and HS3 have the biggest spread in likelihood, but are consistent with the higher coverage samples.

Orkney, we find that they map onto the same ancestral location as the Iron Age samples (Supplementary Fig. 11), confirming that they are of distinct ancestry from the Cornish population and other populations used in building the tree, similarly to the Iron Age samples. As detailed in Supplementary Note 5, our mapping approach crucially depends on an appropriate model for the reference populations. When using the Kent population for building the tree (Fig. 3c), we find that mapping British samples becomes worse (Supplementary Fig. 12), arguably because the Kent population is less genetically defined and more admixed than the group from Cornwall. In such cases we need to model population phylogenies with admixture and gene flow, and further development on rarecoal will enable us to study these more complex scenarios.

Discussion

This study combines large modern sample sets with ancient genomes in a novel way, based on rare allele sharing. On the one hand, the power of rare genetic variants clearly shows the value in whole-genome sequencing of ancient DNA: While SNP capture technology provides a far more economical way to obtain genome-wide data from ancient DNA (ref. 14), it cannot detect rare genetic variants, which as we have shown are necessary to analyse subtle genetic differences between closely related populations. On the other hand, our analysis shows the value of having whole-genome sequence for a large number of modern samples to ascertain rare variants, which fortunately is increasingly becoming the standard for large population scale genetic studies[16–19].

Our analysis of early and middle Anglo-Saxon samples from East England adds significantly to our picture of the Anglo-Saxon period in Britain. In the cemetery at Oakington we see evidence even in the early Anglo-Saxon period for a genetically mixed but culturally Anglo-Saxon community[24,25], in contrast to claims for strong segregation between newcomers and indigenous peoples[7]. The genomes of two sequenced individuals (O1 and O2) are consistent with them being of recent immigrant origin, from a source population close to modern Dutch, one was genetically similar to native Iron Age samples (O4), and the fourth was consistent with being an admixed individual (O3), indicating interbreeding. Despite this, their graves were conspicuously similar, with all four individuals buried in flexed position, and with similar grave furnishing. Interestingly the wealthiest grave, with a large cruciform brooch, belonged to the individual of native British ancestry (O4), and the individual without grave

goods was one of the two genetically 'foreign' ones (O2), an observation consistent with isotope analysis at West Heslerton which suggests that new immigrants were frequently poorer[26,27].

Up to this point we have interpreted the genetic structure of the Anglo-Saxon samples in terms of recent immigrant versus indigenous populations. However, in the absence of a time series through the Romano-British period from the Iron Age to the Anglo-Saxon period, we should also consider the possibility that some of the genetic heterogeneity seen in the Oakington samples arose earlier due to immigration in Romano-British times. We recall that sample O4 lies genetically almost centrally in the Iron Age samples, and O1 and O2 are very close to the later Middle Saxon samples from Hinxton and modern Dutch. For Roman immigration patterns to generate this diverse structure in the fifth to sixth century Oakington samples, one would have to assume strong social segregation with little interbreeding over multiple generations. This seems unlikely given that immigration into Roman-Britain was geographically diverse and consisted of an administrative elite[28] and the military, who would have interbred and recruited locally, particularly in the last decades of the third and fourth centuries[29]. Furthermore, there is no significant Roman settlement at Oakington and no evidence for significant Roman Heritage[30].

Given the mixing apparent $\sim$500 CE, and that the modern population is not more than 40% of Anglo-Saxon ancestry, it is perhaps surprising that the middle Anglo-Saxon individuals from the more dispersed field cemetery in Hinxton look more genetically consistent with unmixed immigrant ancestry. One possibility is that this reflects continued immigration until at least the Middle Saxon period. The unmixed Hinxton group, versus the mixing of the Oakington population, shows that early medieval migration took a variety of forms and that these migrants integrated with the incumbent population in different ways. Full-genome sequences, and new methods such as rarecoal, now allow us to use slight distinctions in genetic ancestry to study such recent events. Further ancient genomes, and methodological improvements to incorporate explicit migration and mixing, will enable us to resolve them in more detail.

Methods

Custom software mentioned here is publically available on www.github.com/stschiff/sequenceTools and www.github.com/stschiff/rarecoal.

DNA extraction. Samples were first treated with UV-light (260 nm) for 20–30 min, and the surface was cleaned with bleach (3.5%) and isopropanol. The sample surface was mechanically removed using a Dremel drill and disposable abrasive discs. Samples were ground to fine powder using a Mikrodismembrator (Sartorius) and stored at 4°C until further use. DNA was extracted in clean room facilities in Adelaide using an in-solution silica-based protocol[31].

Library preparation. Libraries were generated from the Hinxton individuals ($n=6$) with[32] and without enzymatic damage repair (Supplementary Table 1), whereas partial damage repair[33] was performed for the Linton ($n=3$) and Oakington ($n=14$) samples. All 29 libraries were prepared with truncated barcoded Illumina adaptors and amplified with full-length indexed adaptors for sequencing[34]. Protocols evolved over the course of the study with regards to the final library amplification steps. Hinxton DNA libraries were amplified by PCR in quintuplicates for an initial 13 cycles (AmpliTaq Gold, Life Technologies), followed by pooling and purification of the PCR replicates with the Agencourt AMPure XP system. DNA libraries were then re-amplified for another 13 cycles in quintuplicates or sextuplicates, followed by pooling and purification, visual inspection on a 3.5% agarose gel, and final quantification using a NanoDrop 2000c spectrophotometer (FisherScientific). The Oakington and Linton DNA libraries were amplified using isothermal amplifications using the commercial TwistAmp Basic kit (TwistDx Ltd). The amplification followed the manufacturer's recommendations and used 13.4 µl of libraries after the Bst fill-in step, and an incubation time of the isothermal reaction of 40 min at 37 °C, followed by gel electrophoresis and quantification using a Nanodrop spectrophotometer. Following quantification, libraries were re-amplified for seven cycles using full-length 7-mer indexed Illumina primers as described[34], followed by purification with Ampure and quantification using a TapeStation (Agilent).

Library screening. The 23 libraries treated with damage repair were screened for complexity and endogenous DNA on an Illumina MiSeq platform in Harvard in collaboration with David Reich (Supplementary Table 1). When the project started, we had available only the samples from Hinxton, and since all of them had high complexity and high amounts of endogenous DNA (except 12882A, which did not pass screening), we selected all five samples for deep sequencing. We then expanded the project to the other two sites, from which we screened many more samples than we could sequence deeply, so we selected the best four samples (with highest complexity and endogenous DNA) from Oakington and the best from Linton (from which we had fewer samples, and there was only one sample with acceptable complexity for deep sequencing).

Deep sequencing. We first sequenced the five DNA libraries generated from the Hinxton samples in two batches. The first batch consisted of 10 lanes of 75 bp paired end sequencing on an Illumina HiSeq 2500 platform, run in rapid mode. All five samples were multiplexed in this batch. The resulting data was processed (see below) and used to estimate complexity and endogenous DNA to decide further sequencing. The second batch consisted of 42 lanes with similar settings as the first batch, but not multiplexed. Based on the complexity and endogenous DNA estimates, we sequenced sample HI1 and HS3 on 4 lanes each, samples HS1 and HS2 on 8 lanes each and sample HI2 on 16 lanes. In the second batch, we introduced five dark cycles into read 1 to avoid low-complexity issues due to the clean room tags in the library preparation. We also included 5% Phi X sequences to increase the complexity of the first five base pairs of read 2, a common procedure for low-complexity libraries. In case of the samples from Oakington and Linton, we used the protocol used in batch 2 of the Hinxton samples (including dark cycles). We sequenced samples O2 and L on 4 lanes each, sample O4 on 6 lanes, sample O1 on 8 lanes and sample O3 on 10 lanes.

Raw read processing. We filtered out all read pairs that did not carry the correct clean room tags in the first five base pairs of read 2. In case of batch 1 of the Hinxton samples, we also sequenced the clean room tag on read 1, which we also filtered on in these cases. As a second step, we merged all reads searching for a perfect or near perfect overlap allowing at most 1 mismatch between read 1 and the reverse complement of read 2. The merging also took advantage of the fact that we typically had fragments of length 50 pb, which means that many of the 75 bp reads contained the reverse complement of the clean room tag of the other read, and the Illumina adaptors. As a last step, we removed the clean room tags and the adaptors from both ends of the merged reads. Both merging and adaptor trimming was done using a custom programme called filterTrimFastq, available on http://www.github.com/stschiff/sequenceTools.

Alignment. After merging, we ended up with single reads with variable length (on average about 50 bp) for each sample. We aligned those single reads with the programme 'bwa aln'[35] to the human reference, version GRCh37 using the parameter '-l 1024' to turn-off seeding[36]. The alignment was done on a per-lane basis, all alignments were then sorted using 'samtools sort'. For each individual, we then merged the sorted alignments into a single bam file per individual, using 'samtools merge'. We then removed duplicate reads in each alignments using our custom python script 'samMarkDuplicates.py', available also on github. The script checks whether neighbouring reads in the sorted alignments are equal, and removes all but one read if it finds duplicates. Finally, we removed all unmapped reads from the alignments. Despite enzymatic damage repair, some low levels of DNA damage can still be found in the libraries. We used the programme 'mapdamage2' (ref. 37) to measure DNA degradation. For each individual, we first ran mapDamage on chromosome 20 to estimate the degradation profile. For all individuals, the DNA damage profile was found to have an excess of C->T changes at the 5′ end of reads, as expected for ancient DNA, and an excess of G->A changes was found at the 3′ end. However, because the sequencing libraries were treated with UDG, which removes damaged sites in reads, the excess was much lower than in comparable studies without UDG treatment[37].

Mitochondrial and Y chromosome analysis. We called mtDNA and Y chromosome consensus sequences using samtools. Haplogroups were handcurated using public databases (Supplementary Note 2).

Contamination estimates. We estimated possible modern DNA contamination in all ancient samples using two methods. First, we tested for evidence for contaminant mitochondrial DNA[38]. We looked for sites in the mitochondrial genome, at which the ancient sample carried a consensus allele that was rare in the 1,000 Genomes reference panel. We then looked whether there were reads at these sites that carried the majority allele from 1,000 Genomes (Supplementary Note 2). Second, we used the programme 'verifyBamId'[39] to carry out a similar test in the

nuclear genome, again using the 1,000 Genomes reference panel. Contamination estimates are summarized in Supplementary Table 3.

Principal component analysis. We downloaded the Human Origins Data set[13,14] and called genotypes at all sites in this data set for all ancient samples using a similar calling method as described in ref. 14: Of all high-quality reads covering a site, we picked the allele that is supported by the majority of reads, requiring at least two reads supporting the majority allele, otherwise we call a missing genotype. If multiple alleles had the same number of supporting reads, we picked one at random. Principal component analysis was performed using the 'smartpca' programme from EIGENSOFT (ref. 40), by using only the modern samples for defining the principal components and projecting the 10 ancient samples onto these components (Supplementary Fig. 3).

Rare allele sharing analysis. We compiled a reference panel consisting of 433 individuals from Finland ($n=99$), Spain ($n=107$), Italy ($n=107$), Netherlands ($n=100$) and Denmark ($n=20$). The Finnish, Spanish and Italian samples are from the 1,000 Genomes Project (phase 3)[16], the Dutch samples from the GoNL project[17] and the Danish samples from the GenomeDK project[18]. For the Dutch and Danish samples, only allele frequency data was available. In case of the Dutch data set, we downsampled the full data set to obtain the equivalent of 100 samples. All other reference sample variant calls were used as provided by the 1,000 Genomes Project. In addition, we filtered based on a mappability mask[41,42] that is available from www.github.com/stschiff/msmc. We selected all variants up to allele count nine in this reference set and tested for each ancient individual and each of those sites whether the ancient individual carried the rare allele. We called a rare variant (always assumed heterozygous) in the ancient sample if at least two reads supported the rare allele from the reference set. While this calling method will inevitably miss variants in low coverage individuals, the relative numbers of shared alleles with different populations is unbiased.

We accumulated the total number of alleles shared between each ancient sample and each modern reference population, and stratified by allele count in the reference population, up to allele count nine (Supplementary Data 1). We found that sharing with the Dutch and the Spanish population showed the largest variability across the ancient samples. For the plot in Fig. 2a, we divided the sharing count with the Dutch population by the sharing count of the Spanish population for each allele frequency. To plot curves from the Dutch and the Spanish population itself, we sampled haploid individuals from each population by sampling with replacement at every variant site in the reference set. This was necessary because for the Dutch samples no genotype information was publically available, only allele frequency data (Supplementary Note 3).

For the 30 UK10K samples shown in Fig. 2a,b, we started from the read alignment for each individual and called rare variants with respect to the 433 reference individuals in exactly the same way as we did for the ancient samples. For Fig. 2a, the allele sharing counts were then accumulated across the 10 individuals in each group. Error bars for each allele sharing count are based on the square root of each count. For Fig. 2b we added the allele sharing counts between each ancient sample and each reference population up to allele count five, and computed the ratio $NED/(NED+IBS)$, where NED is the sharing count with Dutch, and IBS the sharing count with Spanish (Supplementary Note 3). For the mean and variances shown in Fig. 2b, we excluded outliers as indicated in the caption of the figure. The fraction of Anglo-Saxon derived ancestry is computed for each modern UK10K sample as the relative distance of its relative sharing ratio from the Iron Age mean value compared with the Saxon era mean value, as shown in Fig. 2b, with 0% corresponding to the Iron Age mean, and 100% corresponding to the Anglo-Saxon era mean (Supplementary Note 3, Supplementary Table 4).

Rarecoal analysis. Rarecoal is a new framework to calculate the joint allele frequency spectrum across multiple populations using rare alleles. Given a certain distribution of rare derived alleles across subpopulations (here up to allele count four), and a given number of non-derived alleles, which can be arbitrarily large, we calculate the total probability of that configuration under a demographic model. The model consists of a population tree with constant population sizes in each branch of the tree and split times. To give rise to the data observed in the present, the lineages of the derived alleles must coalesce among each other before they coalesce to any non-derived lineage. We introduce a state space that contains all possible configurations of derived lineages across populations and propagate a probability distribution over this space back in time. Details and mathematical derivations are given in Supplementary Note 6.

We implemented rarecoal in a software package (available from www.github.com/stschiff/rarecoal) that can learn the parameters of a given population tree topology from the data using numerical maximization of the likelihood and subsequent Markov Chain Monte Carlo to get posterior distributions for each split time and branch population size. We did not implement an automated way to learn the tree topology itself, but use a step by step protocol to learn the best topology fitting the data, adding one population at a time (Supplementary Note 5). The outputs from rarecoal are in scaled time. To convert to real time (years) and real population sizes, we used a per-generation mutation rate of 1.25×10^{-8} and a generation time of 29 years.

We tested the method on simulated data using the sequential coalescent with recombination model (SCRM) simulator[43] with the model shown in Fig. 3b with 1,000 haploid samples distributed evenly across the five populations and realistic recombination and mutation parameters. We then learned the model from the European data set as shown in Fig. 3c using an iterative protocol, adding one population at a time and maximizing parameters subsequently to ensure that we are still fitting the right topology (Supplementary Note 5).

For mapping ancient samples on the tree we used the same calling method as in the rare allele sharing analysis. We then added the ancient individual as a separate seventh population to the European tree and evaluated the likelihood for this external branch to merge anywhere on the tree. We restricted the fitting to alleles that were shared with the ancient sample and excluded private variants in the ancient sample, which have high false-positive rates. We also made sure that the age of the ancient sample was correctly modelled into the joint seven-population tree, by 'freezing' the state probabilities from the present up to the point where the ancient sample lived.

For testing the tree-colouring method, we used single individuals from within the reference set and used them as separate sample to be mapped onto the European tree. (Supplementary Note 5).

References

1. Cunliffe, B. *Britain Begins* (Oxford University Press, 2013).
2. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
3. O'Dushlaine, C. T. *et al.* Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* **18,** 1248–1254 (2010).
4. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519,** 309–314 (2015).
5. Weale, M. E., Weiss, D. A., Jager, R. F., Bradman, N. & Thomas, M. G. Y chromosome evidence for Anglo-Saxon mass migration. *Mol. Biol. Evol.* **19,** 1008–1021 (2002).
6. Capelli, C. *et al.* A Y chromosome census of the British Isles. *Curr. Biol.* **13,** 979–984 (2003).
7. Thomas, M. G., Stumpf, M. P. & Harke, H. Evidence for an apartheid-like social structure in early Anglo-Saxon England. *Proc. Biol. Sci.* **273,** 2651–2657 (2006).
8. Topf, A. L., Gilbert, M. T., Dumbacher, J. P. & Hoelzel, A. R. Tracing the phylogeography of human populations in Britain based on 4th-11th century mtDNA genotypes. *Mol. Biol. Evol.* **23,** 152–161 (2006).
9. Busby, G. B. *et al.* The role of recent admixture in forming the contemporary West Eurasian genomic landscape. *Curr. Biol.* **25,** 2518–2526 (2015).
10. Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. *PLoS Biol.* **11,** e1001555 (2013).
11. Zheng, H. X., Yan, S., Qin, Z. D. & Jin, L. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. *Sci. Rep.* **2,** 745 (2012).
12. Balaresque, P. *et al.* A predominantly neolithic origin for European paternal lineages. *PLoS Biol.* **8,** e1000285 (2010).
13. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
14. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).
15. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
16. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
17. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).
18. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6,** 5969 (2015).
19. The UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).
20. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8,** e1002967 (2012).
21. Winney, B. *et al.* People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *Eur. J. Hum. Genet.* **20,** 203–210 (2012).
22. Sajantila, A. *et al.* Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc. Natl Acad. Sci. USA* **93,** 12035–12039 (1996).
23. Sundell, T., Kammonen, J., Halinen, P., Pesonen, P. & Onkamo, P. Archaeology, genetics and a population bottleneck in prehistoric Finland. *Antiquity* **88,** 1132–1147 (2014).
24. Hines, J. The becoming of the English: identity, material culture and language in early Anglo-Saxon England. *Anglo-Saxon Studies Archaeol. Hist.* **7,** 49–59 (1994).

25. Härke, H. Ethnicity, race and migration in mortuary archaeology: an attempt at a short answer. *Anglo-Saxon Studies Archaeol. Hist.* **14,** 12 (2007).
26. Budd, P., Millard, A., Chenery, C., Lucy, S. & Roberts, C. Investigating population movement by stable isotope analysis: a report from Britain. *Antiquity* **78,** 127–141 (2004).
27. Montgomery, J., Evans, J. A., Powlesland, D. & Roberts, C. A. Continuity or colonization in Anglo-Saxon England? Isotope evidence for mobility, subsistence practice, and status at West Heslerton. *Am. J. Phys. Anthropol.* **126,** 123–138 (2005).
28. Eckardt, H., Müldner, G. & Lewis, M. People on the move in Roman Britain. *World Archaeol.* **46,** 534–550 (2014).
29. Petts, D. Military and civilian: reconfiguring the end of Roman Britain in the North. *Eur. J. Archaeol.* **16,** 314–335 (2013).
30. Mortimer, R., Sayer, D. & Wiseman, R. in *Life on the Edge: Social, Political and Religious Frontiers in Early Medieval Europe*. (eds Semple, S., C. Orsini and S. Mui) Neue Studien zur Sachsenforschung 6, in press (Durham UK, 2016).
31. Brotherton, P. *et al.* Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nat. Commun.* **4,** 1764 (2013).
32. Briggs, A. W. & Heyn, P. Preparation of next-generation sequencing libraries from damaged DNA. *Methods Mol. Biol.* **840,** 143–154 (2012).
33. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370,** 20130624 (2015).
34. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** t5448 (2010).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
36. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genom.* **13,** 178 (2012).
37. Jònsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29,** 1682–1684 (2013).
38. Skoglund, P. *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336,** 466–469 (2012).
39. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91,** 839–848 (2012).
40. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
41. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
42. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46,** 919–925 (2014).
43. Staab, P. R., Zhu, S., Metzler, D. & Lunter, G. Scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* **31,** 1680–1682 (2015).

Acknowledgements

We thank everyone who contributed to the archaeological excavations, the sequencing team at the Wellcome Trust Sanger Institute, and David Reich's laboratory for contributing to the characterization of the libraries. We thank Luka Papac for wet lab support at the Australian Centre for Ancient DNA. The Oakington excavations were funded by the Oakington Parish Council, the Institute for Field Research (IFR) and the University of Central Lancashire (UCLan). This work was funded by Australian Research Council grant DP130102158, by the University of Adelaide's Environment Institute and by Wellcome Trust grant 098051.

Author contributions

S.S., A.C., C.T.S. and R.D. designed and oversaw the study. E.P., R.C. and A.L. provided samples from Linton and Hinxton, D.S. provided samples from Oakington. W.H. prepared samples and extracted D.N.A., W.H. and B.L. generated sequencing libraries. S.S., P.P. and R.D. developed methods. S.S., P.P. and R.D. analysed data. E.P., R.C., A.L., L.L., R.M. and D.S. provided archaeological context. S.S., R.D. and D.S. wrote the paper and all contributed comments.

Additional information

Accession codes: The raw sequence data of the 10 samples presented in this paper are deposited at the European Nucleotide Archive (http://www.ebi.ac.uk/ena). The study IDs are ERP003900 (Hinxton samples) and ERP006581 (Oakington and Linton samples).

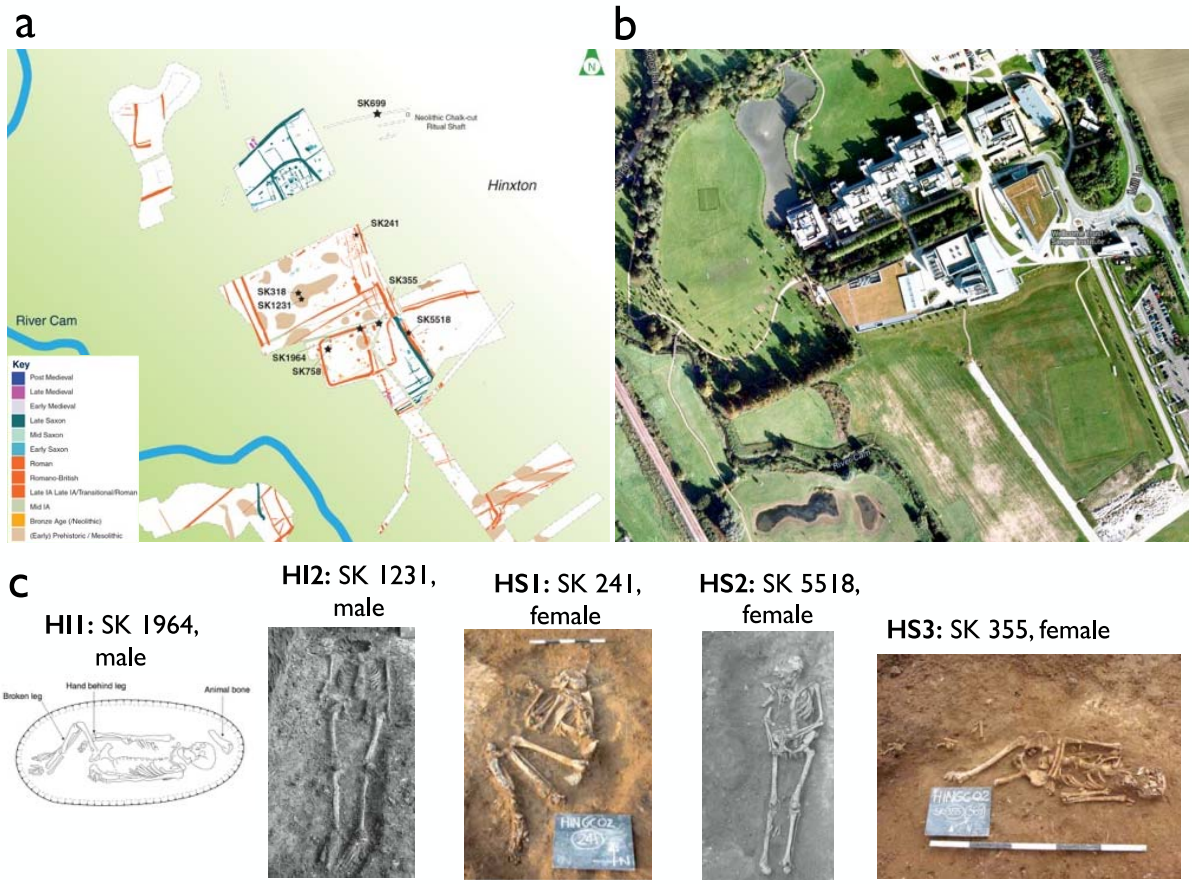
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

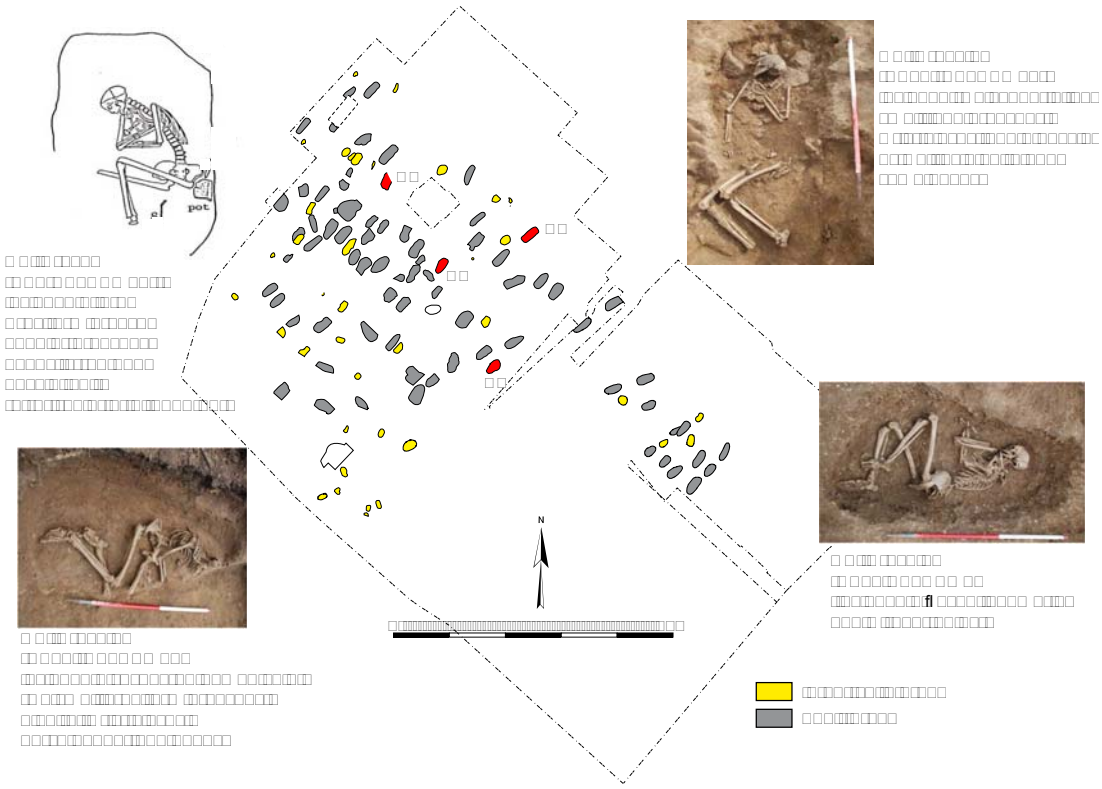
How to cite this article: Schiffels, S. *et al.* Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat. Commun.* 7:10408 doi: 10.1038/ncomms10408 (2016).

Supplementary Figure 1 – Hinxton Site



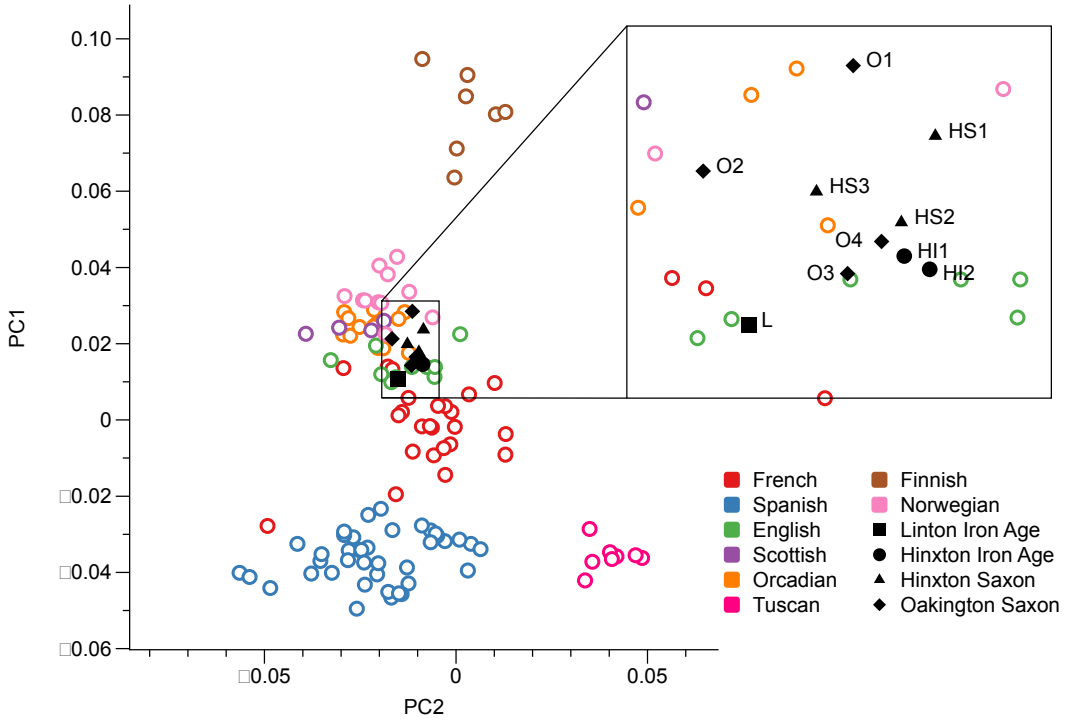
Supplementary Figure 1: Hinxton Site. (a) A plan of the Hinxton archaeological site, with the locations of the skeletal remains. (b) A satellite image of the same area, where today the Wellcome Trust Genome Campus is located. (c) Pictures/Drawing of the 5 samples used in this study.

Supplementary Figure 2 – Oakington Site



Supplementary Figure 2: Oakington Site. A schematic of the early Anglo-Saxon cemetery in Oakington, with graves colored in grey (adult individuals), yellow (infant individuals) and red (the adult individuals used in this study).

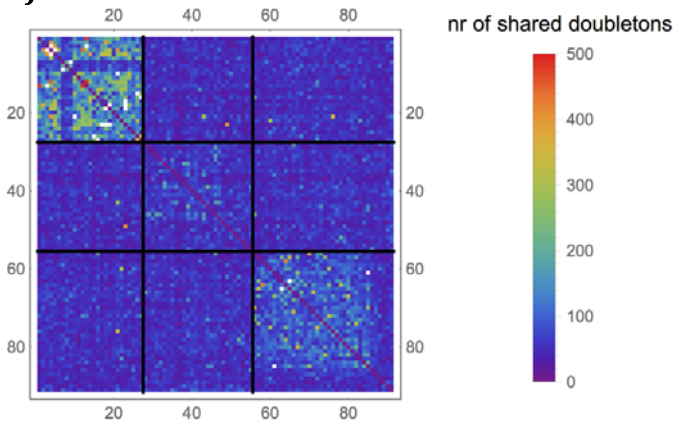
Supplementary Figure 3 - Principal Component Analysis



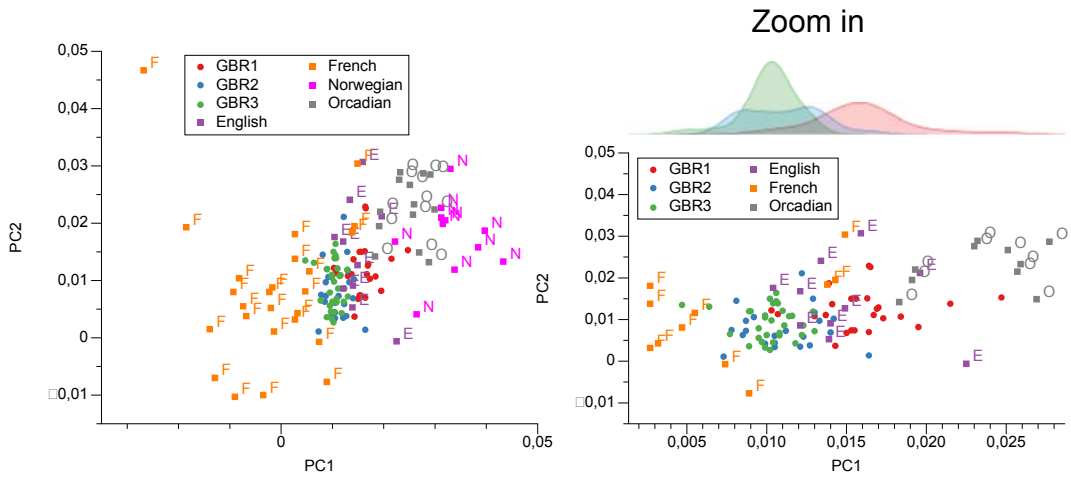
Supplementary Figure 3: Principal component analysis. The first two principal components obtained by analyzing European samples from the Human Origins Data set [10,11] and projecting the ancient samples onto these components. Only populations from Northwestern central Europe are shown. The populations from the Human Origins data set to produce this plot are: Albanian, Bergamo, Bulgarian, Cypriot, Greek, Italian_South, Maltese, Sicilian, Tuscan, English, French, Icelandic, Norwegian, Orcadian, Scottish, Basque, French_South, Spanish, Spanish_North, Belarusian, Croatian, Czech, Estonian, Hungarian, Lithuanian, Ukrainian, Canary_Islanders, Sardinian, Finnish, Mordovian, Russian.

Supplementary Figure 4 – Population Structure in the GBR samples

a)



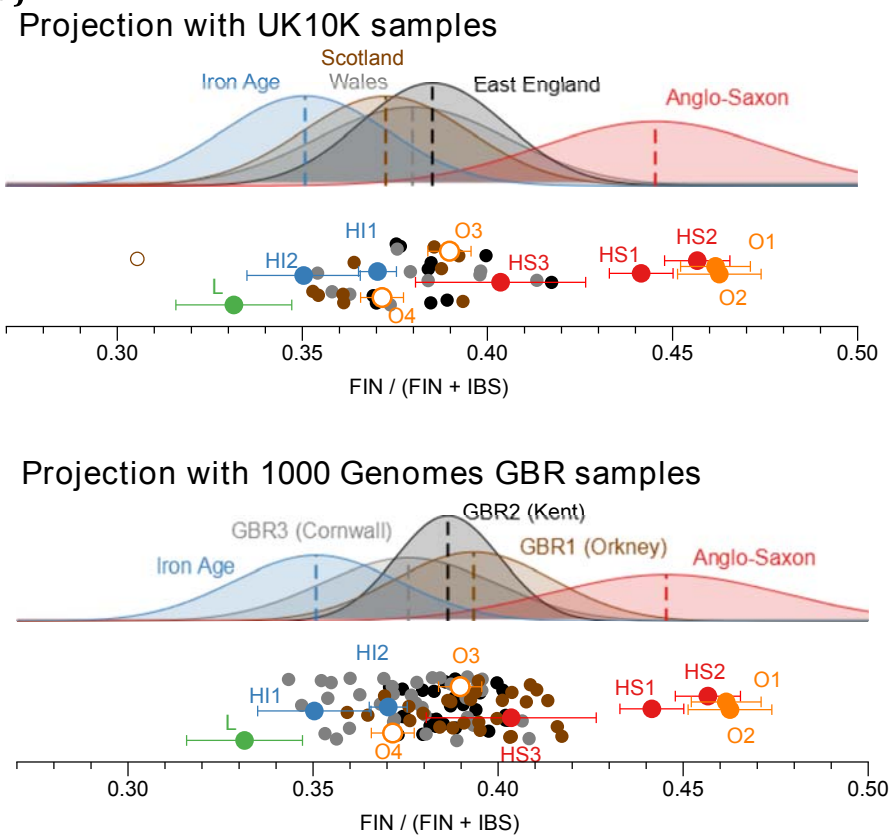
b)



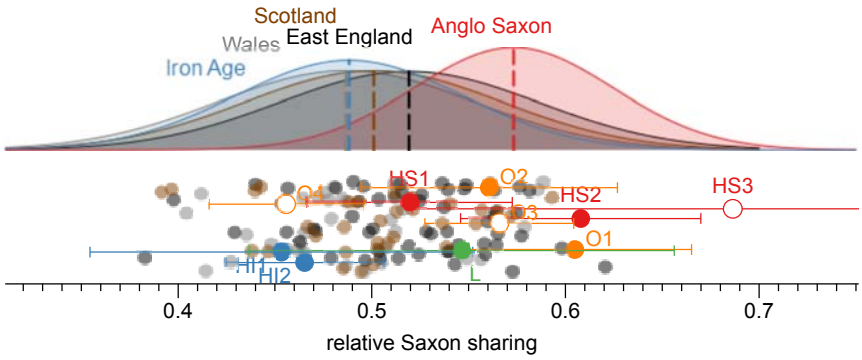
Supplementary Figure 4: Population structure in in the GBR samples. (a) This matrix shows the number of shared doubletons (mutations with allele count 2 within all European 1000 Genomes samples) between two individuals of the 91 GBR samples. The black lines are manually placed to distinguish the three visible clusters. (b) Principal component plot of the 1000 Genomes GBR samples. The three clusters identified in the GBR samples (named GBR1, GBR2 and GBR3) are projected onto selected European samples from the Human Origins data set. We conclude from this analysis that GBR1 corresponds to the Orkney cluster, given its substantially closer location to the Orcadian samples in the PCA plot.

Supplementary Figure 5 – Additional rare variant projections

a)

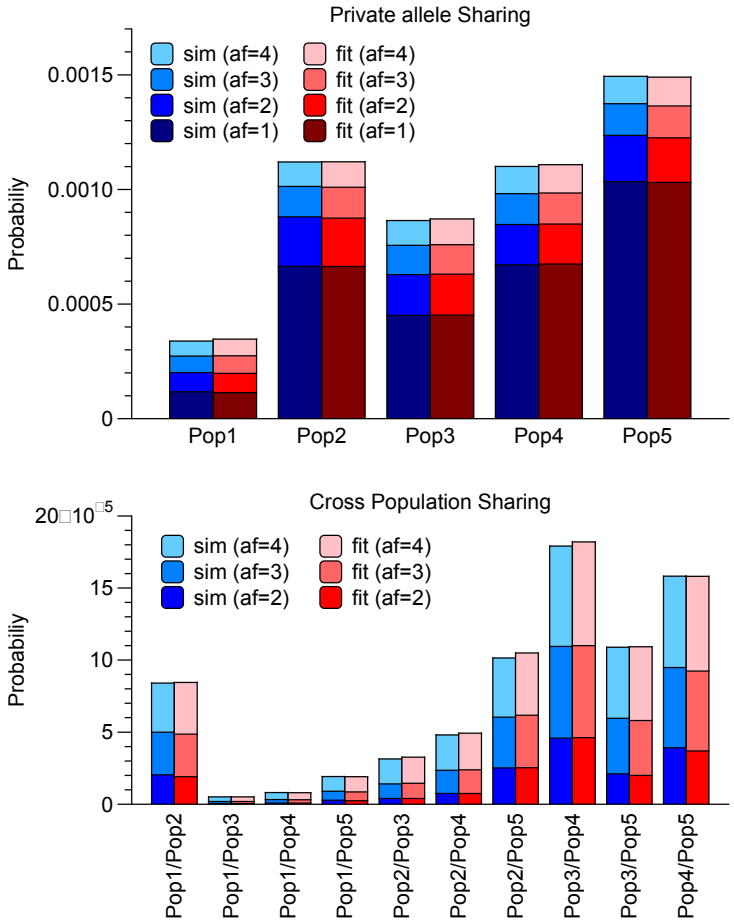


b)



Supplementary Figure 5: Additional rare variant projections. (a) Projection of modern British samples using Finnish vs. Spanish allele sharing, similar to the analysis shown in Figure 2 in the main text and described in Methods, but with the Finnish instead of the Dutch population as an outgroup. The X axes show how many rare variants up to allele count 5 (identified in 433 Europeans) are shared with Finnish samples vs. Spanish samples. The upper plot shows the same modern samples as in Figure 2, from the UK10K project. The lower plot shows 91 modern samples from the GBR population, grouped into three clusters. (b) Allele sharing between UK10K and ancient samples. This figure shows how many rare alleles (identified in 1854 UK10K samples) each UK10K individual from one of the three locations shares with the Anglo-Saxon vs. the Iron Age group (see Supplementary Note 3 for details).

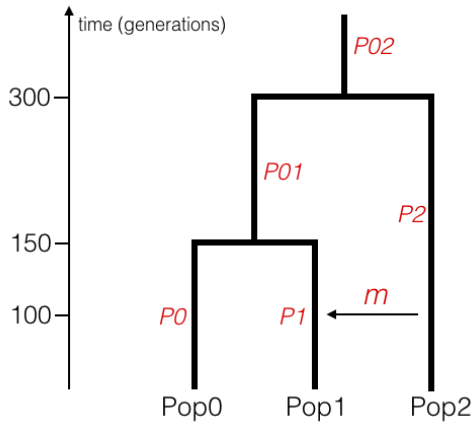
Supplementary Figure 6 – Rarecoal fits of simulated data



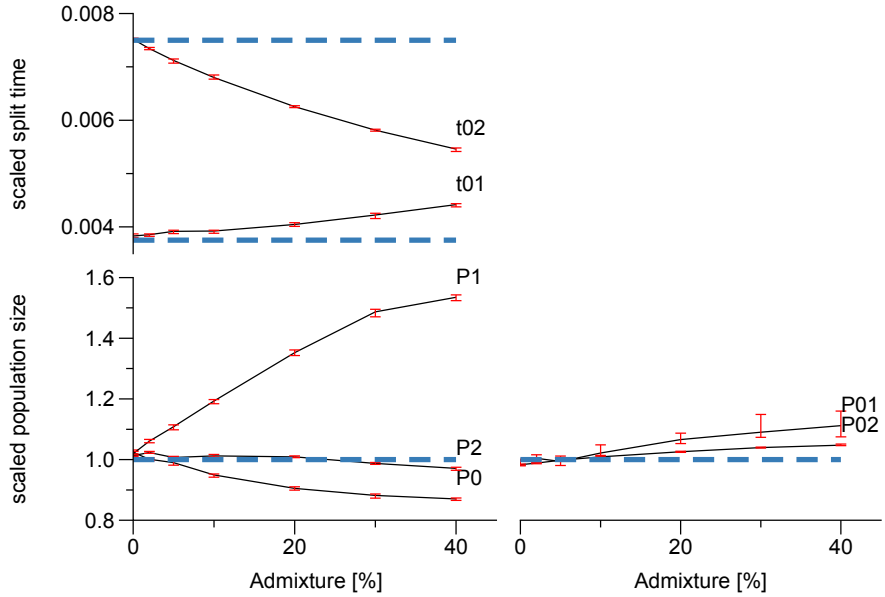
Supplementary Figure 6: Rarecoal fits of simulated data. We compare the theoretical distribution of rare variants predicted by the model estimated in Figure 3b (red) with the true distribution of variants (blue), yielding a good fit of the model given the data. The top panel shows variants private to one population, the lower panel shows variants shared across populations.

Supplementary Figure 7 – Rarecoal estimates under admixture

a)



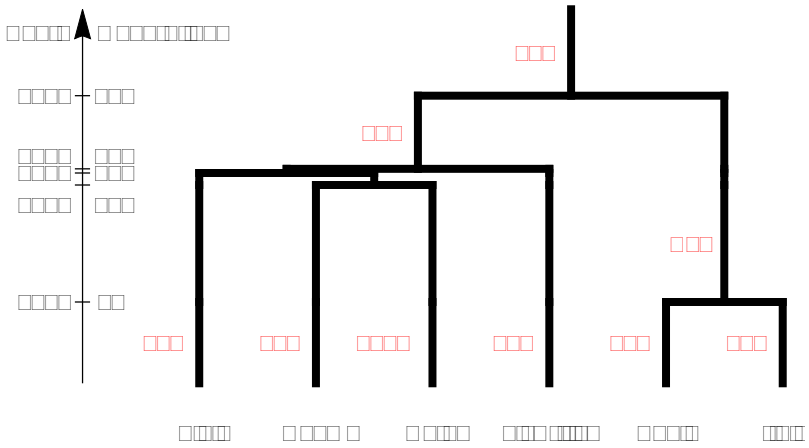
b)



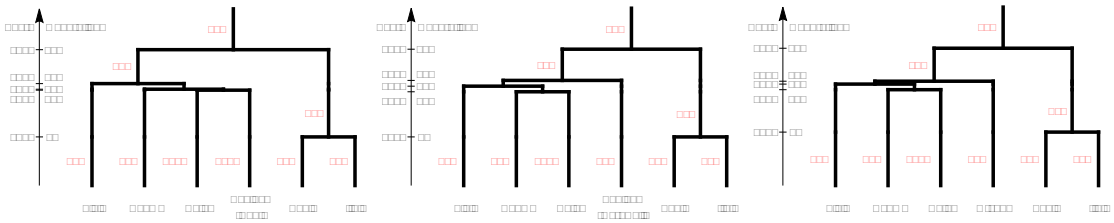
Supplementary Figure 7: Rarecoal estimates of simulations to test robustness under admixture. (a) We simulate three populations, with 100 diploid individuals each, related by two split times, 150 and 300 generations ago. At 100 generations ago, admixture with proportion □ occurs from Pop2 into Pop1. (b) The dashed blue lines indicate the true value, and the x axis denotes the rate of admixture. As can be seen, increasing admixture leads to an increasing deviation of the estimated split times and population sizes from the true parameters.

Supplementary Figure 8 – Maximum likelihood trees of European populations

a)



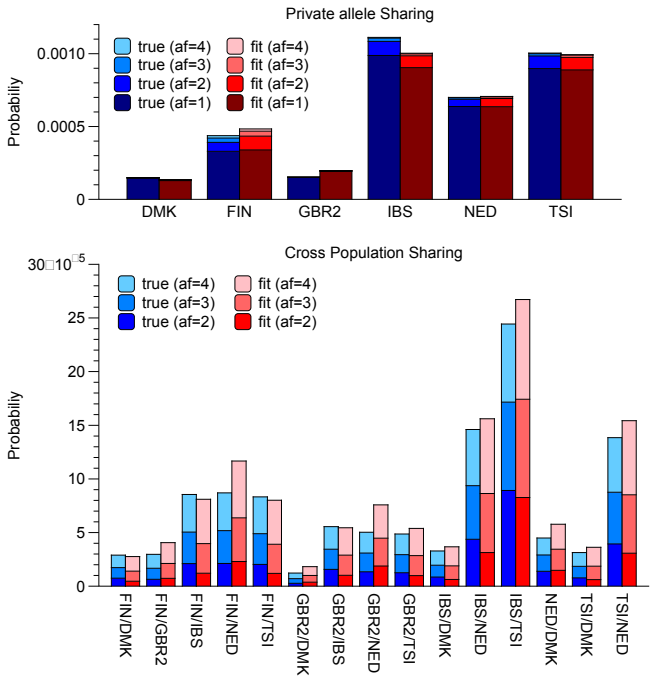
b)



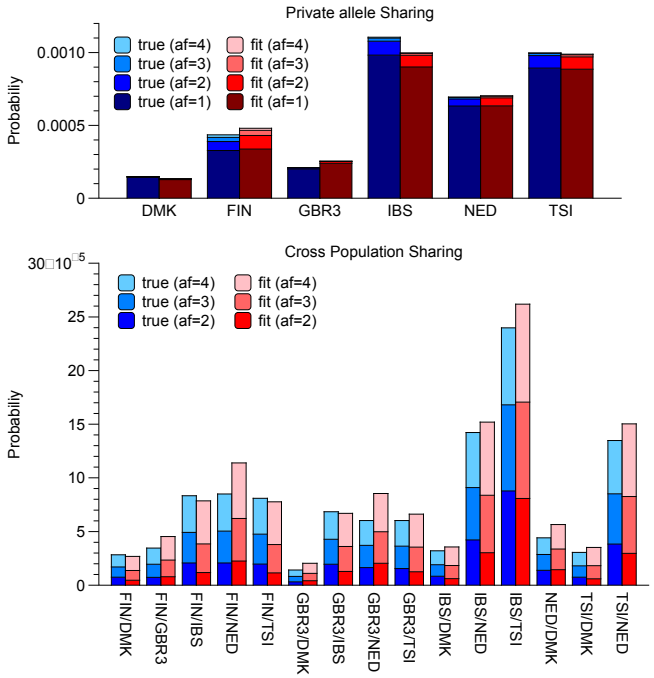
Supplementary Figure 8: Maximum likelihood trees of European populations. (a) European tree estimated from 524 individuals without separating the British samples into subpopulations. Population size estimates are shown in red, split time estimates on the left axis. (b) European trees using the three groups in the GBR sample set separately.

Supplementary Figure 9 – Rarecoal fits of European data

a)

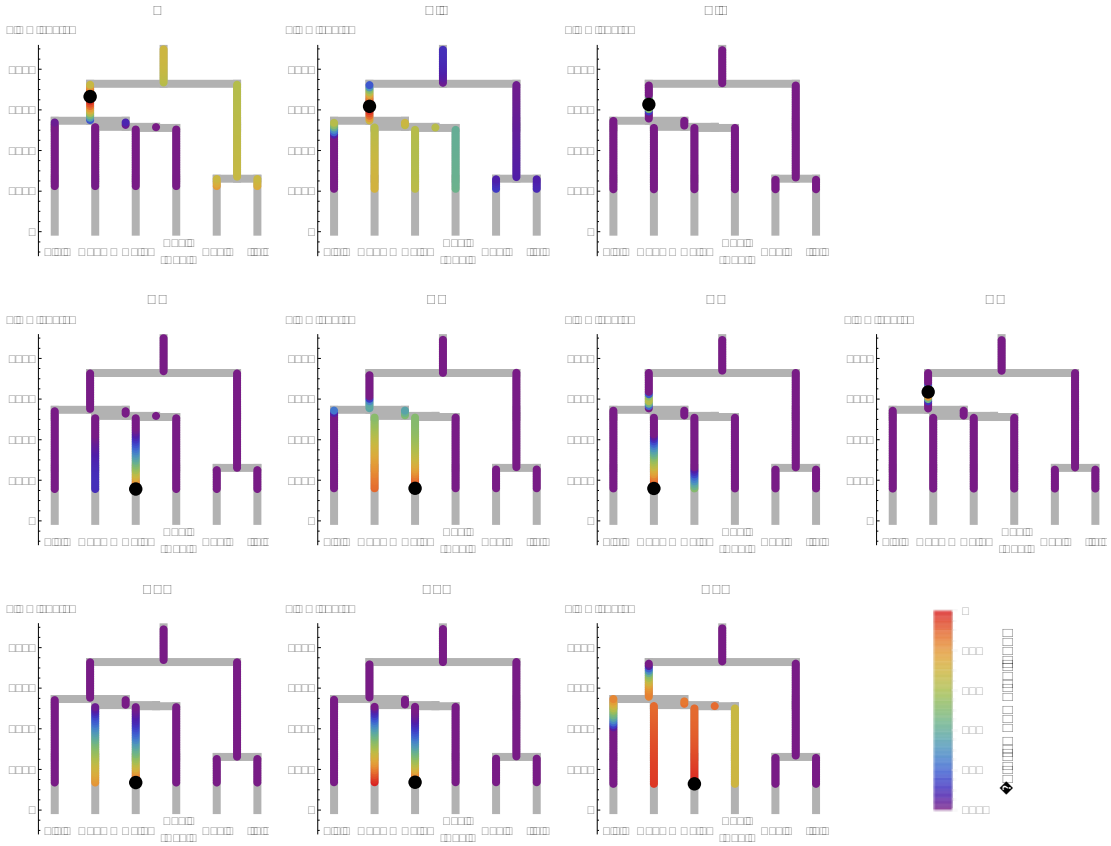


b)



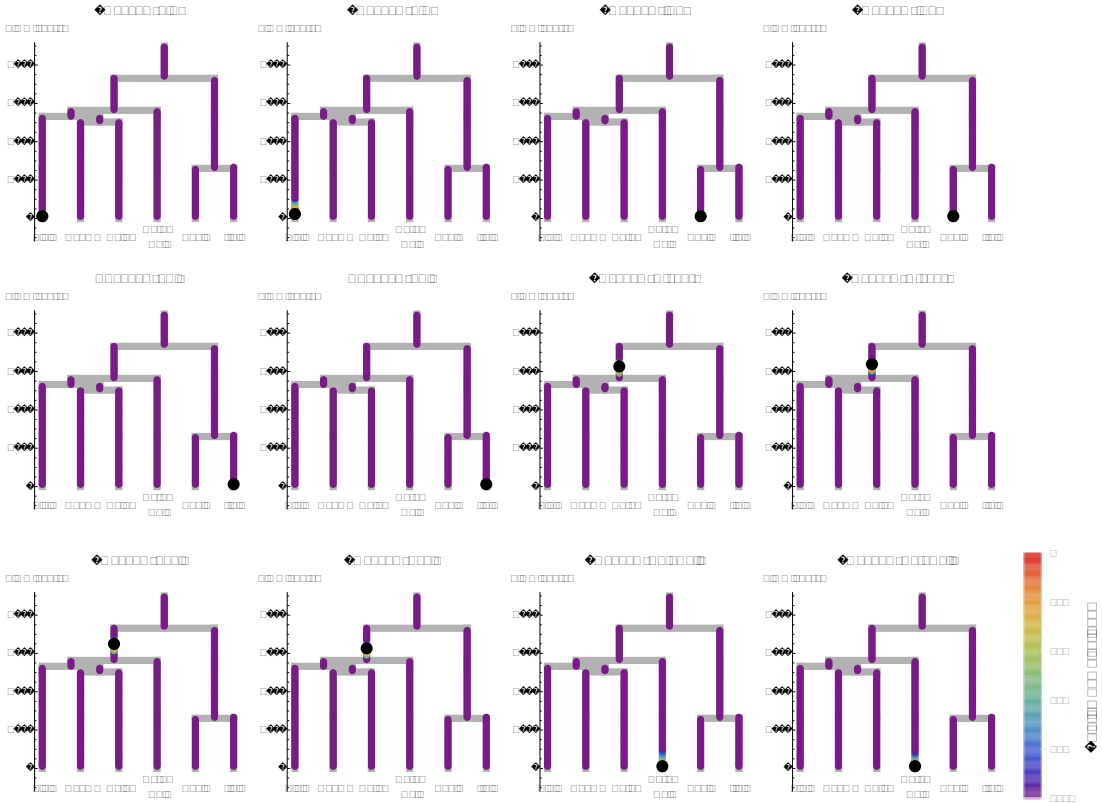
Supplementary Figure 9: Rarecoal fits to European data. Similar to Extended Data Figure 5, we obtain fits between the model obtained on the European samples with the true distribution of rare variants. In a) we fit the tree using samples from Kent (GBR2), as shown in Figure 3c, and in b) we fit the tree shown in Figure 3d, with samples from Cornwall (GBR3). The fit is reasonable, with some systematic differences owing to simplifying assumptions such as constant population sizes and the absence of migration.

Supplementary Figure 10 – Tree mapping using Kent as GBR proxy



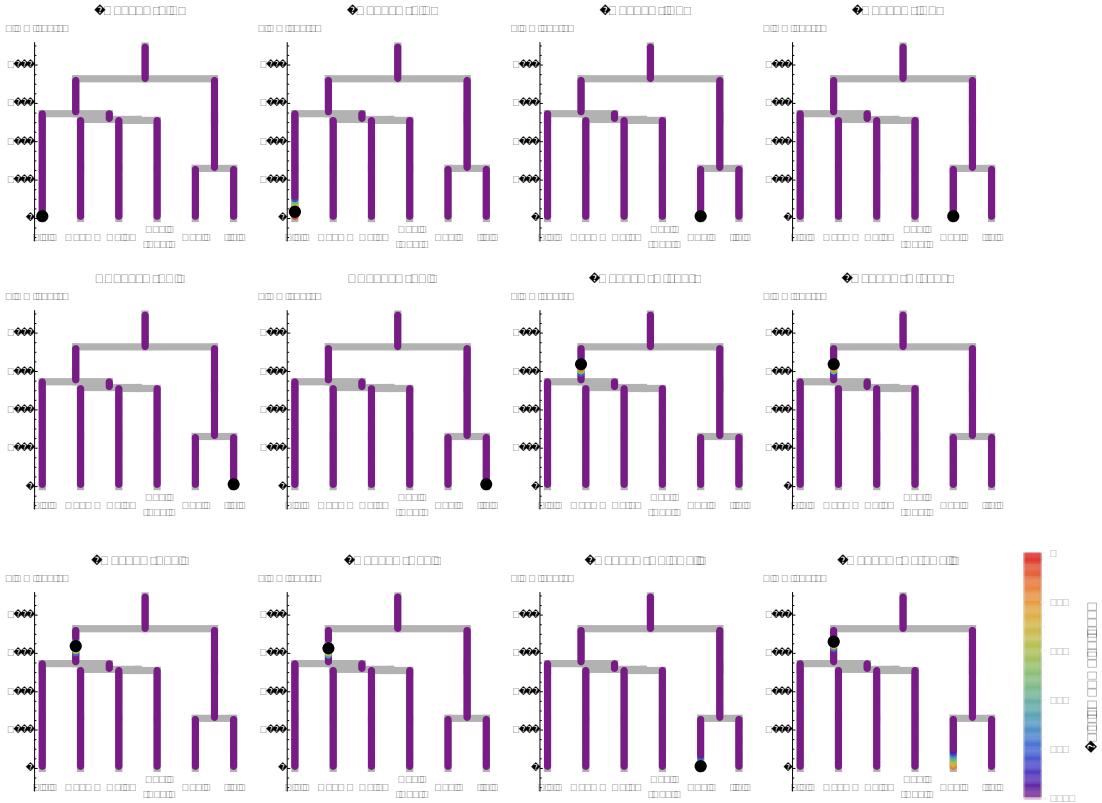
Supplementary Figure 10: Placing ancient samples into the European tree, using the Kent population as British branch. This shows a similar analysis as shown in Figure 4 in the main text (see Supplementary Note 5), but with the Kent population (instead of the Cornwall population) as a proxy for the British branch.

Supplementary Figure 11 – Mapping modern samples onto European tree with Cornwall as British branch



Supplementary Figure 11: Rarecoal tree painting with modern samples. The likelihood surface along the tree (see Supplementary Note 5) for several modern samples from the 1000 Genomes project. Here we used the samples from Cornwall as the proxy for the English population. Most samples map correctly onto the tip of their respective branches, but when we map GBR samples from Kent or Orkney, they map to the Northern European ancestral branch, as expected with an English branch based on Cornwall. The black dot indicates the maximum likelihood merge point onto the tree.

Supplementary Figure 12 - Mapping modern samples onto European tree with Kent as British branch



Supplementary Figure 12: Mapping modern samples from 1000 Genomes into a European tree using Kent as British population branch. A similar figure as Supplementary Figure 11, but with Kent used as the British branch, instead of Cornwall.

Supplementary Table 1 – DNA libraries

Library ID	Sample ID	Article ID	Individual/ museum ID	Sample Type	Site	Repair	Complexity	% endog, DNA
LP26.01	12880A	HI1	SK1964	Second premolar root	Hinxton	USER	n/a	19%
LP26.02	12881A	HS1	SK241	First molar root	Hinxton	USER	n/a	34%
LP26.03	12882A		SK758	Lower first molar root	Hinxton	USER	n/a	n/a
LP26.04	12883A	HS2	SK5518	Upper right canine root	Hinxton	USER	n/a	39%
LP26.05	12884A	HI2	SK1231	Lower third molar root	Hinxton	USER	n/a	85%
LP26.06	12885A	HS3	355	Lower second molar root	Hinxton	USER	n/a	19%
LP49.01	15548A		Grave 57a (1375)	Upper left 2nd molar	Oakington	UDGhalf	0.5	13%
LP49.02	15549A		Grave 59(1395)	Upper left 1st incisor	Oakington	UDGhalf	0.3	55%
LP49.03	15550A		Grave 61(1411)	Lower left 3rd molar	Oakington	UDGhalf	0.9	19%
LP49.04	15553A		grave 66 (1450)	Lower left 3rd molar	Oakington	UDGhalf	7.9	29%
LP49.05	15555A		Grave 78a (1747)	Upper left canine	Oakington	UDGhalf	0.4	41%
LP49.06	15556A		Grave 80 (1740)	Lower left 2nd molar	Oakington	UDGhalf	0.1	1%
LP49.07	15558A	O1	Grave 82 (1779)	Upper left 2nd molar	Oakington	UDGhalf	9.8	76%
LP49.08	15617EBC		extraction blank			UDGhalf		
LP49.09	15560A		Grave 85 (1785)	Upper left 1st premolar	Oakington	UDGhalf	0.4	18%
LP49.10	15568A		Grave 94 (1866)	Upper right 2nd incisor	Oakington	UDGhalf	0.5	54%
LP49.11	15569A	O2	Grave 95 (1870)	Lower right 2nd molar	Oakington	UDGhalf	7.2	89%
LP49.12	15570A	O3	Grave 96 (1882)	Lower left canine	Oakington	UDGhalf	26.6	92%
LP49.13	15575A		GrAVE 112 (2222)	Lower right canine	Oakington	UDGhalf	0.4	13%
LP49.14	15576A		burial 3 (1622)	Lower left 3rd molar	Oakington	UDGhalf	n/a	n/a
LP49.15	15577A	O4	burial7 (1633)	Lower left 3rd molar	Oakington	UDGhalf	100	67%
LP49.16	15618EBC		extraction blank			UDGhalf		
LP50.11	15579A	L	Sk 270	Lower canine	Linton	UDGhalf	4.5	51%
LP50.12	15583A		Sk 352	Upper left 2nd incisor	Linton	UDGhalf	0.1	1%
LP50.13	15586A		Sk 351	Upper right 3rd molar	Linton	UDGhalf	1.4	12%
LP50.14	15589A		Sk 887	Lower right canine	Oakington	UDGhalf	0.3	2%
LP50.16	15683EBC		extraction blank			UDGhalf		

Supplementary Table 1: Library preparation details for all samples that were screened. See Methods for details about library preparation. Only those libraries with labels in column 3 were selected for deep sequencing, based on screening results. Values in the "complexity" columns give the fold coverage of the genome expected after hypothetical sequencing of the entirely library.

Supplementary Table 2 – Radiocarbon dates of samples

Sample	Service reference	Uncalibrated conventional age	2-sigma calibrated age
L	SUERC-14246	2155±35BP	360 - 50 BCE
HI1	OxA-29573	2039 ±27	160 BCE - 26 CE
HI2	Wk-12599	2029±49BP	170 BCE - 80 CE
O1	Beta-397731	1560±30 BP	420 - 570 CE
O2	Beta-397732	1620±30 BP	385 - 535 CE
O3	Beta-397733	1600±30 BP	395 - 540 CE
O4	Beta-397734	1590±30 BP	400 - 545 CE
HS1	OxA-29573	1288 ±25	666 - 770 CE
HS2	OxA-X-2565-12	1320± 45	631 - 776 CE
HS3	OxA-29572	1230 ±25	690 - 881 CE

Supplementary Table 2: Radiocarbon dates of samples. The table gives the uncalibrated and calibrated C14-dates for all sequenced samples. The reference starting with SUERC is from the [illegible] The reference starting with Wk is from the [illegible] The references starting with OxA are from the [illegible] The references starting with Beta are from [illegible] Calibrated dates are computed from the Oxcal computer program (v4.2) of C. Bronk Ramsey, using the 'IntCal13' dataset.

Supplementary Table 3 – Contamination estimates

Sample	mtDNA Coverage	Informative Sites	N_{cons}	N_{alt}	mtDNA estimate	Nuclear estimate
L	78	0			n/a	0.00012
HI1	1145	4	5341	3	0.00056	0.00005
HI2	2177	1	2473	13	0.0052	0.00887
O1	642	3	9168	0	<0.00033	0.01495
O2	652	0			n/a	0.01219
O3	410	6	35913	4	0.00011	0.01312
O4	255	0			n/a	0.01505
HS1	1020	4	4197	3	0.00071	0.01090
HS2	537	7 (6)	3290 (2673)	89 (10)	0.027 (0.0037)	0.01018
HS3	587	6	4206	0	<0.00071	0.00009

Supplementary Table 3: **Contamination estimates.** DNA contamination estimates based on mitochondrial and nuclear DNA. Numbers are contamination fractions on a 0-1 scale. For O2, O4 and L, no mtDNA estimate could be generated because there were no informative sites. The relatively high contamination estimate of HS2 is due to a single site in the hypervariable region, which could reflect natural heteroplasmy. The estimate without that site is given in parentheses for that individual. See Supplementary Note 2 for details.

Supplementary Table 4 – Estimates of the Anglo-Saxon ancestry fraction in modern Britain

Data set	Group	With HS3	Outgroups	Anglo-Saxons	Anglo-Saxons StdDev	Iron Age	Iron Age StdDev	Value for Group	StdDev for Group	Fraction	StdDev
UK10K	East	Yes	Dutch,Spanish	0,607	0,015	0,504	0,026	0,543	0,013	**0,38**	**0,21**
UK10K	Wales	Yes	Dutch,Spanish	0,607	0,015	0,504	0,026	0,535	0,013	0,30	0,22
UK10K	Scotland	Yes	Dutch,Spanish	0,607	0,015	0,504	0,026	0,536	0,013	0,31	0,22
UK10K	East	No	Dutch,Spanish	0,614	0,007	0,504	0,026	0,543	0,013	0,35	0,19
UK10K	Wales	No	Dutch,Spanish	0,614	0,007	0,504	0,026	0,535	0,013	0,28	0,21
UK10K	Scotland	No	Dutch,Spanish	0,614	0,007	0,504	0,026	0,536	0,013	0,29	0,21
UK10K	East	Yes	Finnish,Spanish	0,445	0,02	0,351	0,016	0,385	0,014	**0,36**	**0,20**
UK10K	Wales	Yes	Finnish,Spanish	0,445	0,02	0,351	0,016	0,380	0,018	0,31	0,23
UK10K	Scotland	Yes	Finnish,Spanish	0,445	0,02	0,351	0,016	0,372	0,016	0,22	0,22
1000G	Kent	Yes	Finnish,Spanish	0,445	0,02	0,351	0,016	0,386	0,01	**0,37**	**0,17**
1000G	Cornwall	Yes	Finnish,Spanish	0,445	0,02	0,351	0,016	0,376	0,016	0,27	0,22
1000G	Orkney	Yes	Finnish,Spanish	0,445	0,02	0,351	0,016	0,393	0,015	0,45	0,21
UK10K	East	No	Finnish,Spanish	0,456	0,008	0,351	0,016	0,385	0,014	0,32	0,17
UK10K	Wales	No	Finnish,Spanish	0,456	0,008	0,351	0,016	0,380	0,018	0,28	0,20
UK10K	Scotland	No	Finnish,Spanish	0,456	0,008	0,351	0,016	0,372	0,016	0,20	0,20
1000G	Kent	No	Finnish,Spanish	0,456	0,008	0,351	0,016	0,386	0,01	0,33	0,14
1000G	Cornwall	No	Finnish,Spanish	0,456	0,008	0,351	0,016	0,376	0,016	0,24	0,19
1000G	Orkney	No	Finnish,Spanish	0,456	0,008	0,351	0,016	0,393	0,015	0,40	0,17

Supplementary Table 4: Estimates of the Anglo-Saxon ancestry fraction in modern Britain. Estimates of the Anglo-Saxon component in the modern British population, using different outgroup populations (Dutch and Finnish vs. Spanish) and different British populations as test cases. We include both the case with and without HS3 as a member of the Anglo-Saxon group. O3 and O4 are always excluded because they seem admixed or of non-Anglo-Saxon ancestry (see Figure 2 in the main text). The three estimates including HS3 for the East of England or Kent are highlighted. Details on how the values in this list are computed can be found in Supplementary Note 3.

Supplementary Note 1 – Archaeological sites and sample descriptions

Linton Site

Between 2004 and 2010 investigations by Oxford Archaeology East (funded by Cambridgeshire County Council) on land at Linton Village College, Cambridgeshire (NGR TL 55547 46984), produced evidence of over four and a half thousand years of human activity. The 8ha site lies in an agriculturally rich area on the lower valley slopes of the River Granta, just outside the village of Linton. A range of features and deposits of later Neolithic to post-medieval date was revealed across most of the areas investigated. These included a series of later Neolithic Grooved ware pits, two ring-ditches (remains of burial mounds), a Middle to Late Bronze Age enclosure and later Iron Age settlement evidence; the latter associated with an inhumation and metalworking debris of the same date. Roman features included a field system and trackway, in addition to the remains of a possible animal-powered mill and a number of neonate burials. Post-Roman activity was represented by an Early Saxon enclosure, five Middle Saxon inhumations (a possible execution cemetery) and a quantity of 17th-century items possibly related to a documented Civil War skirmish.

Analysed sample from Linton

Linton Skeleton 270 (AKA 2270) (sample L in the main Text):

A poorly-preserved contracted ('crouched') inhumation of a female aged over 50 in a shallow, oval grave (1.1m x 0.7m) located in proximity to an area of settlement-related features. The burial was aligned north to south, and the skeleton was laid on its right side, with the head facing west. Analysis of the skeleton revealed that the individual was 1.58m (+/- 4.3 cm) tall. Osteoarthritis and spondylosis deformans were present in her spine and wrist, while enamel hypoplasia indicates that she experienced health stress during childhood.

Additional samples (Anglo-Saxon) from Linton

Linton Skeletons 351 and 352:

A group of three graves containing five skeletons was uncovered in the area of a former Roman trackway. One of the graves, aligned north-east to south-west, contained three individuals (sks 350, 351 and 352) that were all apparently buried during a single event. The grave was sub-rectangular, with steeply sloping sides and a flat base, and measured 1.91m long, 0.92m wide and 0.20m deep.

The initial burial appears to have been that of an older child of around 12 years of age (sk 352), who had been positioned along the eastern side of the grave in a supine position with the head to the south-west. Some pathological changes were noted on this skeleton including evidence for growth arrest, metabolic disease (cribra orbitalia and porotic hyperostosis) and mild trauma. No evidence for peri-mortem injuries was observed. This burial was followed by the interment of a child of around five years of age (sk 350) that was placed in the south-west corner of the grave.

The final burial was that of a mature adult female, aged over 45 (sk 351), who had been placed centrally in the grave on top of skeletons 352 and 350. This

individual had been decapitated prior to burial and the head had been deposited within the grave first. The skeleton was in a loosely extended, supine position with the feet to the south and right femur lying over the top of the skull. Both arms were flexed at the elbows, with the left arm lying across the torso and the right angled outwards 'akimbo' from the body. Several pathological conditions were observed, including developmental anomalies, maxillary sinusitis, Schmorl's nodes and joint disease. Peri-mortem sharp-force trauma, associated with head removal, was present on the fourth and fifth cervical vertebrae.

Hinxton Site

Extensive archaeological investigations were undertaken in Hinxton, South Cambridgeshire by Oxford Archaeology East between 1993 and 2014 on behalf of the Wellcome Trust [1,2]. The investigations, which centred around Hinxton Hall and the Genome Campus, extended on either side of the River Cam and were set within a rich archaeological landscape (Supplementary Figure 1). The ancient course of the Icknield Way crosses the site, which itself lies 1.5 kilometres north of the Roman town at Great Chesterford. This post-glacial valley landscape attracted humans to hunt and make flint tools from the Late Upper Palaeolithic (□ 10,000 BC) and into the Mesolithic and Early Neolithic periods until eventually the first tree clearances to enable farming and more permanent settlement began. This area also became a focus for more ceremonial activities associated with the dead during both the Middle Bronze Age and the Iron Age to Roman periods, represented by burials and a mortuary enclosure. From the Middle Iron Age until the Middle Romano-British period the site appears to have been in continuous agrarian use, specialising in animal husbandry, until its apparent abandonment.

The land was not resettled until the Early to Middle Saxon period when activity included a small scatter of timber houses and sunken-featured buildings and associated features. By the Late Saxon period, settlement had coalesced in the northern part of the site (Hinxton Hall), associated with an ordered field system. During the 11th century a large ditch enclosed the settlement, and several new timber buildings were constructed. This may have been the documented Hengest's Farm, which gave modern Hinxton its name. Further Late Saxon discoveries were made in Ickleton, on the western side of the River Cam, where a working area probably associated with flax retting and wood working was found. To the south of the main enclosed settlement were the remains of a small hamlet, also occupied during the Saxo-Norman and earlier medieval period and seemingly abandoned by the early 13th century. A number of Anglo-Saxon burials were scattered around the eastern limits of the settlements, buried within silted up ditches and pools and within an isolated grave.

Analysed samples from Hinxton

□□□□□□□□□□□□, sample HI1 in the main text:

Skeleton 1964 was that of an old male, buried supine with its legs extended, within a grave located in the north-east corner of the mortuary enclosure. Analysis indicates that this skeleton was dolichocranic, or had a relatively long skull, and had maxillary sinusitis, vertebral disc herniation (Schmorl's nodes) and an oblique fracture of the right lower leg that had healed. At 159.0 cm tall, the individual was within the normal range for the period. Dental pathology was

observed indicating that the individual had periodontal disease, advanced caries, abscesses and had also lost all of their molars and lower right second premolar before death.

□□□□□□□1231 (sample HI2 in the main text):

An isolated burial placed within an infilled pond that had also previously contained a Bronze Age skeleton. The Late Iron Age/Early Roman skeleton was that of a middle/old adult male who had been placed in a north-east to south-west orientated grave in an extended, supine position with their arms by their side and their head in the north-east. Their stature was 174.1cm. They had lost a number of teeth prior to death and the skeleton also displayed evidence of caries and abscesses. In addition to showing evidence of joint disease (osteoarthritis), Schmorl's nodes, maxillary sinusitis and metabolic disease (cribra orbitalia), some pathological changes were observed may have been caused by repetitive activity involving the shoulder from a young age.

□□□□□□□241, sample HS1 in the main text:

Buried within a shallow oval grave cut into the top of a major boundary ditch, skeleton 241 was that of a middle aged/old female placed in a crouched position. This individual measured 158.6 cm in stature. Ante mortem tooth loss had affected the two lower mesial incisors only, and this unusual position may indicate that an occupational use of the teeth, or perhaps trauma, had resulted in their loss. Other dental conditions included caries and periodontitis. Osteoarthritis was present on some joints, while evidence of Schmorl's nodes and metabolic disease (cribra orbitalia) was also observed.

□□□□□□□ □□□□□sample HS2 in the main text:
A very large sub-oval grave or pit lay to the south of that containing sk 241, and was also cut into the boundary ditch: it contained the skeleton of a middle aged/old female (50+) that was in a supine position. This individual measured 153.6 cm in stature and had suffered ante mortem tooth loss, caries and abscesses; evidence of trauma, Schmorl's nodes non-specific bone inflammation and joint (including osteoarthritis) and metabolic disease were also present.

□□□□□□□355, sample HS3 in the main text:
A grave located adjacent to the entrance way of an enclosure contained the skeleton of a young/middle adult female. Buried in a supine position with her legs flexed, the skeleton was aligned roughly north to south with the arms lying across the abdomen. This individual had an estimated stature of 163.5 cm and showed evidence of Schmorl's nodes and trauma, including a healed fracture on the right arm.

Additional samples from Hinxton

Skeleton 758 (Middle to Late Iron Age)

Skeleton 758 was an adolescent (less than 16 years) of unknown sex buried within the north-east corner of the mortuary enclosure, where it had been inserted into the top of an existing pit. The individual was buried supine with the legs extended and arms by their sides. Schmorl's nodes were present on the spine.

Oakington Site

Early Anglo-Saxon Cemeteries

Furnished Anglo-Saxon burials have been studied for nearly three centuries, based on radiocarbon dates and artistic styles we know that these equipped graves date between the late fifth and early eighth centuries [3]. The earliest phase of burial rituals dates to the fifth and sixth centuries and have been referred to as Migration Period, Pagan or early Anglo-Saxon graves [4]. These cemeteries are predominantly found in the south and east of England from Dorset to Northumberland with regional variation evident within the burial rite [5]. Grave goods include weapons, for example; spears, swords or shield bosses. Grave goods might also be dress objects, for example; brooches, beads, pins or buckles. Also included are containers, parts of animals or Roman artefacts curated and deposited hundreds of years after their manufacture, for example; spoons, coins or rings and brooches. Grave furnishings like these vary according to male or female gender and with age [6]. Many graves have no surviving artefacts at all, and we can only speculate about the organic furnishings which may have been present.

In the early 20th century archaeological interpretations attributed these graves to specific Historical narratives, for example, Anglo-Saxon migration or invasion events. More recent interpretations, however, do not consider funerals to have been the product of static cultural processes, but dynamic and mutable interactions during which communities and individuals expressed and constructed their own identities [7-9]. Participants at these events were associates with different backgrounds including, but not limited to; extended families, households, kinship groups, dependents (slaves and/or children) and social elites depending on who the deceased was. Each burial event was unique and each one was specific to and contingent upon a particular historical moment meaningful to the community that created it.

Oakington early Anglo-Saxon Cemetery

Oakington is a small village in Cambridgeshire, UK, seven kilometres northwest of Cambridge. It was named □□□□□□□□□and □□□□□□□□□□in the Domesday Book of AD 1086 (VCH 1989:192-195). The Oakington early Anglo-Saxon cemetery was first identified in 1926 when three burials were found as a result of cultivation [10]. The site (Supplementary Figure 2) was rediscovered in 1993 during the construction of a children's playground and in 1994 the Cambridge County Council's Archaeological Field Unit excavated an area of 140 sq. m, identifying 24 human skeletons [11]. In 2000 the 1993-94 skeletons were interred within a brick lined vault to the west of the excavated area. In 2006 and 2007 the same archaeological group, then known as CAMARC, excavated a further area of 450 sq. m ahead of the construction of the village's new Recreation Centre, the excavators recorded 17 skeletons. Between 2010 and 2015 the cemetery was systematically excavated by a University of Central Lancashire team (UCLan), with support from Oxford Archaeology East (OAE, formerly CAMARC) and with outreach activities organised by members of staff from Manchester Metropolitan University [12].

By the end of the final excavation season in 2014, a total of 128 individuals had been excavated from an area of approximately 1800 sq. m. Radiocarbon dates

from the skeletal remains and the artefacts from within the graves provide a primarily sixth century date for the cemetery. Preliminary skeletal investigations show that 34 individuals were female, 25 male, 7 adults remain unidentified, 27 individuals were sub-adults aged between 6 and 12, and 35 were below the age of 5. This unusually high number of younger individuals may identify Oakington as a central place in a regional kinship network [13]. The artefacts from the 2010-2014 excavations are currently being conserved and the skeletal remains are being analysed for publication.

Samples used in this study

Oakington [OAKQUW93/11] 1633 Grave 1 (O4 in the main text) was the first grave excavated in 1993 during the playground development, she was a female in her 'mid 40s' and was 1.61m or 5'3" tall [11]. The body was positioned on her right hand side with the head to the south west of the grave facing down towards the knees. She was buried facing east and positioned with her legs flexed forward and arms crossed at her chest. The grave was furnished with a large cruciform brooch, a pair of wrist clasps, a pair of annular brooches, 14 amber beads, two blue beads, a silver coloured glass bead and a large pot sherd. She was also found with a strap-end, knife and a D shaped iron buckle. In 2000 the skeleton was buried in a vault adjacent to the cemetery site. This vault was excavated by the UCLan team in 2012 and the 1633 remains were found stored within labelled containers.

Oakington [OAKQUW12] 1779 (O1 in the main text) was in grave 82 and was excavated in 2012 by the UCLan team. The grave contained the remains of an adult female laid with her head to the south of the grave and facing east. She was positioned on her back with legs slightly flexed to the right. Her left arm crossed over the torso and was placed over the right chest area. The grave was furnished with two copper alloy small long brooches, a pair of wrist clasps, a buckle, a knife and some beads. Preservation within this grave is mixed, the skull is in good condition but the lower part of the body and pelvis was missing, probably as a result of burrowing.

Oakington [OAKQUW12] 1870 (O2 in the main text) was in grave 95 and was excavated in 2012 by the UCLan team. The grave contained the remains of an adult female laid with her head to the south and facing east. She was positioned on her right hand side with legs flexed forward and crossed. Her arms were placed out in front and her left arm was flexed at the elbow to position her hand under her chin. This grave was not furnished with objects.

Oakington [OAKQUW12] 1882 (O3 in the main text) was in grave 96 and excavated in 2012 by the UCLan team. An adult female laid with her head to the south and facing west. The body was placed on the left hand side with legs crossed and slightly flexed, her arms and hands were positioned to the front. The grave was furnished and included two small copper alloy cruciform brooches, a knife, wrist-clasps, purse hanger, two beads and a perforated copper disc, which may have been a Roman coin. The skeleton was truncated by the construction of the playground and was missing parts of the right tibia and fibula, sections of both radius and ulna and a portion of the skull.

Other Graves Sampled

Oakington Sk887 [OAKQUW07] grave 40. An adult female buried supine with her head to the south. Her left leg was flexed placing her foot under the right leg below the knee. Her right arm was flexed and her hand was placed on the abdomen area. The grave was furnished with 77 amber and glass beads, a pair of wrist clasps, two small copper alloy cruciform brooch, a Roman finger ring, an iron buckle and an iron knife.

Oakington [OAKQUW11] 1375 grave 57a. An adult female aged between 25 and 30 years, she was buried supine with her lower left arm flexed to place her hand over the abdomen area. The grave was furnished with a cruciform brooch and two small long brooches, 21 amber beads, 4 glass beads, wrist clasps, belt fittings and an iron knife. The woman in grave 57 had a foetus across her pelvic cavity, this foetus lay low and transverse suggesting an obstetric problem such as shoulder presentation, and was probably the cause of this double fatality [4].

Oakington [OAKQUW11] 1395 grave 59. An adult female buried flexed on her right side with her head to the south and facing east. Her arms were placed in front of her and crossed over, her left arm was placed on the left knee. This grave was furnished with two copper alloy small long brooches, glass beads and wrist clasps.

Oakington [OAKQUW11] 1411 grave 61. An adult female buried supine with her head to the south and facing east, it appears to be slumped forward over her chest. She was buried with two decorated gilt saucer brooches of a Cambridgeshire type, wrist clasps, an iron knife and an iron purse ring.

Oakington [OAKQUW11] 1450 grave 66. An adult female buried supine with her legs crossed and her lower right arm placed over the stomach area. Her head was to the south and faced west. She was buried with a complete pottery vessel to the south of the grave placed by the head. She had a number of amber beads and two pierced copper alloy pendants. She was also buried with two trefoil small long brooches, a pair of wrist clasps, a copper alloy pin, and iron key/latch lifter belt hanging set and a Roman spoon. She had a large pottery fragment at her feet.

Oakington 1740 [OAKQUW11] grave 80. An adult female buried in a semi flexed position on her right hand side head to the south and facing east. Her right elbow was placed in front, and her hand reached back to clasp a set of beads at her chest. Her left arm was flexed at the elbow. Her lower legs were truncated by the 1993/4 excavation. The grave was furnished with 46 amber beads and 22 glass beads in at least two strings, she had two small silvered disc brooches, strap end, wrist clasps, and an iron girdle hanger which included an iron ring, latch lifters and a copper alloy chatelaine. She was also found buried with a fully articulated bovine.

Oakington [OAKQUW12] 1866 grave 94. An adult male [?] buried supine with the head to the south and slumped onto the chest. His left arm was flexed at the elbow and his hand was placed over his chest. His right leg was flexed over the left at the knee crossing the right leg twice. The grave was furnished with a knife.

Oakington [OAKQUW12] 1785 grave 85. An adult female, buried in a flexed position to the left with her head to the south. Her right arm was placed over the

abdomen. The grave was furnished with a bone comb, an iron ring and an iron knife.

Oakington [OAKQUW12] 1747 grave 78a. An adult female buried in a double grave alongside a child. The adult was buried prone with the head to the south and face down. Her legs were crossed and may have been tied. Her right arm passes under her body and the right hand was positioned to clasp a collection of beads and a brooch by the left side of the head. Her left arm passes under her body and her fingers were resting on the child's left arm. The adult was furnished with 17 glass beads, wrist clasps, a small long brooch, an iron knife and an animal bone.

Oakington [OAKQUW13] 2222 grave 112. An adult [?] skeleton buried supine with the head to the south and facing east. The spine curved to the east and both arms were slightly flexed with both hands over the pelvis. The grave was furnished with a knife between the hands and the pelvis.

Acknowledgments for Oakington excavations

Dr Ran Boytner, the Institute of Field Research and the Heritage Lottery Fund for supporting the project. Dr Allison Jones and Dr Gary Bond (UCLan) for funding and administrative support. Dr Faye Sayer and Alison Draper (Manchester Metropolitan University) for outreach and conservation work. Dr Ash Lenton (Australia National University) for on-site surveying. For supervision work during the excavations we would like to thank: Dr Rob Wiseman (Oxford Archaeology East), Sam D Dickinson, Allison Card, Rick Sayer, Meredith Carroll, Vicki Le Quelenec, Tracy Shuttleworth, Kie Leeming, Clare Bedford, Caitlin Halton, Alex Batey, James Hodgsen, Debbie Sale and Justine Biddle.

Supplementary Note 2 - Mitochondrial DNA and Y chromosome analysis

The haplotyping was done by calling consensus sequences using samtools 0.1.19 and bcftools version 1.1, with "samtools mpileup -u -t DPR -r MT", and "bcftools view -v snps". This lists snps that differ from the reference rCRS, which belongs to haplogroup H2a2a1. The haplogrouping was handcurated using the phylotree build 16 from www.phylotree.org [14]. There were a few private snps, as is to be expected from ancient samples, see the table below. We also note that the sample HS3 was a perfect match with the rCRS, apart from one indel.

The haplogroups (listed the following table) are among the most common modern haplogroups in the UK. The haplogroup H1 is found in 13,83% of modern 1000 genomes GBR samples, H in 20,21%, T in 11.43%, K1 in 1.06% and U5 13,83%, [15]. Approximate times for haplogroups can be inferred from [16], and based on these, the age of the haplogroups of our samples are between 8,501 years and 1,428 years with large error margins. The ages of each individual haplogroup is consistent with the radiocarbon dating of the samples.

Individual	MT Haplogroup	Private SNP positions	Age of haplogroup [years]
HI1	K1a1b1b	195	4471
HS1	H2a2b1	72, 195	2088
HS2	K1a4a1a2b		2276
HI2	H1ag1	152	2312
HS3	H2a2a1		2094
O1	U5a2a1	150	2636
O2	H1g1		1901
O3	T2a1a	7941	2165
O4	H1at1		2935
L	H1e	14110, 16362	2026

Several previous studies have associated the haplogroup U5 with hunter-gatherer origins, and the haplogroups H,T, K1 as having Neolithic origins in Europe, see [17] and references therein.

Y chromosome haplogroups

The Y haplogroups were called by first calling Y chromosome genotypes using "samtools mpileup -u -r Y -f ". The coverage of the HI1 sample was very low on the Y chromosome, and therefore we restricted our attention to the unique regions within the male-specific part of the Y chromosome reference sequence, that spanned 8.97 Mb in nine separate regions[18]. The Supplementary Table S1 in Wei, et al. [18] was used to filter our Y chromosome calls in HI1 and HI2. We did not do any further filtering, in the hope of capturing at least a few diagnostic SNPs. We compared the informative SNPs to the ISOGG database (http://www.isogg.org/tree/), and determined that the haplogroup of HI1 is R1b1a2a1a2c, and the haplogroup of HI2 is R1b1a2a1a2c1.

The coverage on HI1 on the diagnostic sites is 1x up to 3x, using a minimum mapping quality of 37. We have 7 derived alleles and 7 ancestral alleles.
If we exclude the sites where the allelic state is T or A in the transition polymorphism, we have two markers (L21, S461) left supporting the haplogroup R1b1a2a1a2c, so we conclude that HI1 was probably in haplogroup R1b.

For HI2, the coverage ranges from 1x to 14x on the diagnostic sites is with the mapping quality of 19 and above. We have 15 ancestral alleles and 13 derived alleles. If we exclude the sites where the allelic state is T or A in the transition polymorphism and require mapping quality of at least 30, we have markers D1857, P241, CTS3575, L21, S245, S461. These markers point to the haplogroup R1b1a2a1a2c. It is therefore possible that both HI1 and HI2 could be in the same haplogroup. HI2 has the marker M269, while there is no coverage on HI1 on that site. The incidence of haplogroup R1b1a2 (R1b-M269) is 78.1% in Cornwall, 62.0% in Leicestershire, and 92.3% in Wales. [19]. In the 1000 genomes GBR cohort, 34 out of 46 male samples belong to haplogroup R1b1a2 making it the most common haplogroup in the UK with 73.9% incidence. Both R1b1a2a1a2c (HI1), and R1b1a2a1a2c1 (HI2) are found once in the GBR of 1000 genomes [20].

The following table lists lists the diagnostic genotype calls for HI1:

SNP/marker	Position	Haplogroup	Ref	Alt	Call
PF6454,CTS2664	14416216	R1b1a2	G	A	?
P257,PF2950,U6	14432928	G	A	G	-
PF5896, P244	14433100	P1	G	A	?
PF2952,S314,U2	14577177	G	A	G	-
PF6541,L52	14641193	R1b1a2a1a	C	T	?
F1857,P337,PF5901	14898094	P1	A	G	+
L269,PF3135	14958218	G	C	T	-
PF2955,L116,S284	14989721	G	G	C	+
L402	15204708	G	G	T	-
U21	15204710	G	C	A	-
L21,M529,S145	15654428	R1b1a2a1a2c	C	G	+
S492	16720013	R1b (investigation)	T	C	-
S245,Z245	22200784	R1b (investigation)	C	G	-
S461,Z290	28632468	R1b1a2a1a2c	G	C	+

The inference column contains a ``-'' for the ancestral allele, a ``+'' for the derived allele, and a ``?'' for a derived allele which could be due to post-mortem damage.

The calls for HI2 are:

SNP/Marker	Position	Haplogroup	Ref	Alt	Call
CTS241, DF13,S521	2836431	R1b1a2a1a2c1,	A	C	+
S144, L20	14231292	R1b1a2a1a2b1a1	A	G	-
PF6454, CTS2664	14416216	R1b1a2	G	A	?
U23	14423856	G	A	G	-
P257, PF2950, U6	14432928	G	A	G	-

P244, PF5896	14433100	P1	G	A	?
L382, M3523, PF2951	14469411	G	A	C	-
F1794	14522828	R1b1a2	G	A	?
S314,PF2952, U2	14577177	G	A	G	-
P240, PF5897	14598808	P1	T	C	?
U12	14639427	G	C	A	-
L52, PF6541	14641193	R1b1a2a1a	C	T	?
L32, PF3266,S148,U8	14692227	G2a2b	C	T	-
D1857,P337,PF5901	14898094	P1	A	G	+
L116, PF2955, S284	14989721	G	G	C	-
PF2956, U3	14993358	G	G	A	-
P241, M173	15026424	R1	A	C	+
PF2957,M201	15027529	G	T	G	-
CTS3575	15037433	R1b1a2	C	G	+
PF2958	15086183	G	C	G	-
L402	15204708	G	G	T	-
U21	15204710	G	C	A	-
PF3134, U33	15275200	G	G	C	-
L21,M529,S145	15654428	R1b1a2a1a2c	C	G	+
S492,Z384	16720013	R1b (investigation)	T	C	-
Z2542,CTS8221	17885577	R1b1a2a1a2c1,	C	T	?
S245,Z245	22200784	R1b (investigation)	C	G	+
S461,Z290	28632468	R1b1a2a1a2c	G	C	+

Contamination Estimates

Contamination estimates using the mitochondrial DNA were done using a comparison against the 1000 genomes database. We identified private or near-private consensus alleles in each individual, requiring the minor allele frequency to be less than 5% in the 1000 genomes cohort of modern DNA. We required the quality score to be at least 50, but did not put a restriction on coverage, since coverage was very high to start with. Furthermore, we excluded the positions where either C or G was the consensus allele, because there is a chance that these are due to post-mortem misincorporations. We did a point estimate of mtDNA contamination following Skoglund, et al. [21]. We assumed independence of the bases, and estimated

$$\hat{\square} \square \frac{\square_{\square\square\square}}{\square_{\square\square\square\square} \square \square_{\square\square\square}}$$

If no alternative allele was found, the upper confidence limit was calculated as the value of c at P=0.05 in the binomial distribution

$$\binom{\square}{\square} \square^{\square} \square\square \square \square\square^{\square\square\square}$$

where $\square \square \square$ and $\square \square \square_{\square\square\square}$. In the cases where no diagnostic sites were found, the contamination could not be estimated. Estimates are listed in Supplementary

Table 3. The comparably high contamination level of HS2 is based one site, 16245, where there are 617 calls supporting T and 79 calls supporting C. HS2* has been calculated by removing this one site. The site 16245 is in the D-loop, or hypervariable region of mitochondrial DNA and it is possible that allele counts on this site are within the natural variation of heteroplasmy. We note that in 1000 genomes cohort there are 10T and 1064C. In addition to the estimates from MT DNA, we used a program called "verifyBamId" [22], which estimates autosomal contamination using the 1000 Genomes reference panel.

Supplementary Note 3 – Rare allele sharing analysis

The main processing for rare allele sharing is described in the Methods section of the paper. Here we provide some additional analysis that we performed to replicate the main results.

Relative allele sharing using Finnish and Spanish outgroups

In addition to the UK10K samples shown in Figure 2 of the main text, we performed a similar analysis using the GBR samples from the 1000 Genomes project. As described in Supplementary Note 4, we identify three subpopulations in the GBR samples, which we can conclusively identify with samples from Cornwall, Kent and the Orkney Islands. In this analysis, we could not use the Dutch and Spanish populations as an outgroup, because the GBR genotypes were called jointly with the Spanish samples from the 1000 Genomes project, while the Dutch samples were called indepently. Therefore, using the Dutch and Spanish populations as outgroups would result in biases towards allele sharing with the Spanish samples. Therefore, we use the Finnish samples from the 1000 Genomes project as outgroup.

Supplementary Figure 5a shows two projections of modern British samples using the Finnish and Spanish populations as outgroup. In the first, we used the same individuals from the UK10K project as used in Figure 2. It shows that the choice of the outgroup (Dutch vs. Finnish) has little influence on our estimate of Anglo-Saxon ancestry in the East of England. In both cases, the samples from the East of England and Kent, respectively, place at about 40% between the Iron Age and the Anglo-Saxon samples. Expectedly, in this projection using the Finnish outgroup, the samples from the Orkney islands share substantially more rare alleles with the Finnish than do the other groups from the GBR samples (Kent and Cornwall) and all three groups from the UK10K project.

Projecting UK10K samples directly onto ancient samples

While the results shown Figure 2 in the main text and in Supplementary Figure 5a above are based on allele sharing with outgroup populations, we also tried a more direct approach of comparing allele sharing with Anglo-Saxon vs. Iron Age samples. Here we took the entire TwinsUK data set from UK10K (with genotype calls provided by UK10K, cite), consisting of 1854 individuals from across the UK, as a reference panel and computed allele sharing of each ancient sample with subpopulations from Wales, East England and Scotland, using all variants up to allele count 37 (1%) in the full data set. In this case, because we had to normalize out coverage differences between the ancient samples, we divided the sharing counts for each ancient sample by the number of shared variants with TwinsUK with allele counts 37 through 370 (1%-10%). We then computed for each TwinsUK sample the mean normalized sharing count with the Iron Age group (H1, H2 and L) and with the Anglo-Saxon era group (HS1, HS2, O1 and O2). We did the same calculation for each ancient individual, by first removing that individual from the two groups above and comparing to the rest of each group. We include samples O3 and O4 for comparison, but they were not used to compute the mean and standard deviation shown in the red Gaussian curve (Supplementary Figure 5b).

We do not try to estimate an Anglo-Saxon component from this analysis because the noise is much stronger than the signal, but we note that the results here are qualitatively consistent with the analyses using outgroups, in particular with the East English samples being somewhat closer to the Anglo-Saxon samples than the groups from Wales and Scotland.

Estimating the Anglo-Saxon component in modern England

Supplementary Table 4 summarizes our estimates of the Anglo-Saxon component in the modern British population, using different outgroup populations (Dutch and Finnish vs. Spanish) and different British populations as test cases. We include both the case with and without HS3 as a member of the Anglo-Saxon group. O3 and O4 are always excluded because they seem admixed or of non-Anglo-Saxon ancestry (see Figure 2 in the main text).

The values and standard deviations in Supplementary Table 4 are the relative sharing fraction of the group indicated in column 2, using the outgroups indicated in column 4. The second-last column gives the estimate of the Anglo-Saxon component in that group using the simple formula

$$\text{[illegible]} = \frac{\square\ \square\ \square}{\square\ \square\ \square}$$

Where $\square$ is the value of the modern-British group (e.g. from Kent), $\square$ is the value for the Iron Age group, and $\square$ is the value for the Anglo-Saxon group. The standard deviation of the fraction is computed using the standard error propagation:

$$\text{[illegible]} = \sqrt{\left(\frac{\square\square}{\square\ \square\ \square}\right)^{\square} \square\ \square\square^{\square} \left(\frac{\square\ \square\ \square}{\square\square\ \square\ \square\square^{\square}}\right)^{\square} \square \left(\frac{\square\square\square\square\ \square\ \square\square}{\square\square\ \square\ \square\square^{\square}}\right)^{\square}}$$

We obtain very consistent results for the South and East of England (highlighted in Supplementary Table 4), using different outgroups and different sample sets. The 1000 Genomes group from Kent and the UK10K samples from the East of England have on average an Anglo-Saxon component of 38% or 37% respectively, with a large spread of up to 21%, which reflects variability among the samples. Samples from Cornwall and Wales have consistent results around 30%, again with a large spread. The Scottish samples from UK10K, in contrast have a similarl Anglo-Saxon component as Wales when using the Dutch outgroup, but a lower component when using the Finnish outgroup. We believe that the result using the Dutch outgroup is appropriate, given that it most strongly separates Anglo-Saxon from Iron Age samples. When excluding sample HS3 from the Anglo-Saxon group, this group gets more defined and further away from the modern and Iron Age samples, resulting in a lower estimate of the Anglo-Saxon component, of around 32-35% in East and Kent samples, depending on the outgroup.

Supplementary Note 4 – Population substructure in the GBR samples from the 1000 Genomes Project

The GBR samples from the 1000 Genomes Project [20] were collected from three sites: Kent, Cornwall and the Orkney Islands. We counted doubleton mutations, i.e. mutations with allele count 2, shared by only two GBR individuals, and generated a count matrix for all pairs of samples (Supplementary Figure 4a). The three subpopulations generate a visible pattern in shared doubletons. The matrix shows that the GBR samples are ordered with respect to sampling location and that they fall into three distinct clusters of 27, 28 and 36 individuals, respectively. In particular the first and third cluster exhibits notable excess allele sharing within the cluster, reflecting relatively strong genetic drift in comparison with the second cluster.

We selected the overlap of SNPs with the Human Origins data set [23,24], and generated a PCA plot of all GBR samples projected onto selected European samples (Supplementary Figure 4b). The PCA shows that the first cluster (GBR1) falls more closely with the Orcadian samples from the Human Origins data set than the other two clusters, so we conclude that the first cluster contains the samples from the Orkney Islands.

From the PCA we cannot infer which of the second and third cluster is sampled from Kent and which from Cornwall. A recent publication on British population structure [25] shows that the population from Cornwall is relatively drifted and internally well defined, which suggests that GBR3 is Cornwall and GBR2 is Kent. Furthermore, as shown in Supplementary Note 5, we used rarecoal to find the best fitting phylogeny of 5 European populations plus each of the three GBR clusters separately, and find that the second cluster forms a clade with the Dutch population, while the third cluster forms an outgroup to the rest of Northern Europe. Given the known Anglo-Saxon influence from the Dutch and German coast into the South East of England, we conclude that the second cluster contains the samples from Kent, and the third cluster contains the samples from Cornwall.

Supplementary Note 5 – Rarecoal Analysis

Rarecoal program

The rarecoal method (Supplementary Note 6) is implemented in a command line tool called "rarecoal", and available on https://github.com/stschiff/rarecoal. This command has several subcommands that are documented in detail on the github-webpage, and of which the following are relevant to this analysis:

- "rarecoal maxl": This command finds the maximum likelihood estimates for all parameters specified in a model. This tool performs a greedy search using the Nelder-Mead-Simplex optimization method and should only be used to get a preliminary estimate of the maximum.
- "rarecoal mcmc": This command performs a Markov-Chain Monte-Carlo simulation on the likelihood function to find the local optimium and get posterior distribution confidence intervals for each parameter. This program will automatically perform a burnin phase which will take as long as needed to find the local maximum, and then perform 1000 MCMC iterations to obtain the confidence intervals
- "rarecoal find": This command takes an additional population or sample and tries every possible place on an existing tree for that additional branch to merge onto the tree. It will output the maximum likelihood branch point.

All outputs of the programs are scaled. To get real times in generations, scaled times need to be multiplied by □□□, and to get real population sizes, scaled population sizes should be multiplied with □□. In our case, □□ □ □□□□□□.

Testing Rarecoal with simulated data

We defined a simple population-tree, as shown in Figure 3b of the paper. We used the SCRM simulator [26] with the following command line to simulate 20 chromosomes of 100Mb:

```
scrm 1000 1 -l 100000 -t 100000 -r 80000 100000000 -I 5
200 200 200 200 200 -ej 0.00125 2 1 -ej 0.0025 4 3 -ej
0.00375 5 3 -ej 0.005 3 1 -en 0.00000001 1 0.1 -en
0.00000002 2 2.0 -en 0.00000003 3 1.0 -en 0.00000004 4
5.0 -en 0.00000005 5 10.0 -en 0.00125001 1 1.0 -en
0.0025001 3 0.5 -en 0.00375001 3 0.8 -en 0.005001 1 1.0
```

The tree topology of this tree is (((0, 1), ((2, 3)), 4)), with branches ordered left to right as in Figure 3b in the main text. We first obtained maximum likelihood estimates of only the split times, and a globally fixed population size. Note: all times are scaled with □□□ (not □□□ as in the command line above), and all population sizes are scaled by □□.
This first round of maximization using "rarecoal maxl" is summarized in the following table:

Parameter	True value	Initial value	Estimate
□(□□□)	0.0025	0.001	0.00271
□(□□□)	0.005	0.002	0.00242

□□(□□□) □□□	0.0075	0.003	0.00452
□□(□□□□)□(□□□) □□□	0.01	0.004	0.00592
□ □□□□□□	1	1	0.859

We then used these estimates as starting point for the full model optimization, with separate population size estimates in each internal and leaf-branch of the tree. We denote the population size parameters with N, using as subscript the subtree of the node below that branch. The results are summarized in the following table, including confidence intervals for each parameter as obtained by "rarecoal mcmc":

Parameter	True Value	Median Estimate	95% CI
□(□□□)	0.0025	0.002790	(0.002773, 0.00284)
□(□□□)	0.005	0.005078	(0.00506, 0.00510)
□□(□□□) □□□	0.0075	0.00779	(0.00776, 0.0078)
□□(□□□□)□(□□□) □□□	0.01	0.00979	(0.00973, 0.00982)
□ □	0.1	0.1055	(0.1052, 0.1057)
□ □	2	2.38	(2.35, 2.42)
□ □	1	1.006	(1.003, 1.01)
□ □	5	5.08	(5.03, 5.14)
□ □	10	10.60	(10.48, 10.73)
□ (□□□)	1	0.90	(0.89, 0.91)
□ (□□□)	0.5	0.52	(0.51,0.53)
□ □(□□□) □□□	0.8	0.64	(0.63, 0.65)
□ □(□□□□)□(□□□) □□□	1	0.98	(0.97, 0.99)

Simulating a lower sample size

In the real data, we have diploid sample sizes of about 100 for the Finnish, British, Spanish, Italian and Dutch samples, and only 20 for the Danish population. To see whether the lower sample size in the Danish population creates a bias on the estimates, we generated a simulation similar to the one above, but with only 20 samples for the last population. The command line was

```
scrm 940 1 -l 100000 -t 100000 -r 80000 100000000 -I 5
200 200 200 200 40 -ej 0.00125 2 1 -ej 0.0025 4 3 -ej
0.00375 5 3 -ej 0.005 3 1 -en 0.00000001 1 0.1 -en
0.00000002 2 2.0 -en 0.00000003 3 1.0 -en 0.00000004 4
5.0 -en 0.00000005 5 10.0 -en 0.00125001 1 1.0 -en
0.0025001 3 0.5 -en 0.00375001 3 0.8 -en 0.005001 1 1.0
```

The MCMC analysis on this dataset was started from the same values as in the analysis of the full simulation, and yielded the following results:

Parameter	True Value	Median Estimate	95% CI
□(□□□)	0.0025	0.00279	(0.00277, 0.00280)
□(□□□)	0.005	0.005	(0.00598, 0.00503)
□□(□□□)□□□	0.0075	0.00814	(0.00809, 0.00816)
□□(□□□□)□(□□□)□□□	0.01	0.00961	(0.00958, 0.00966)
□□	0.1	0.106	(0.105, 0.106)
□□	2	2.42	(2.38, 2.45)
□□	1	0.994	(0.98, 1.00)
□□	5	4.89	(4.84, 4.94)
□□	10	11.7	(11.0, 12.3)
□(□□□)	1	0.87	(0.86, 0.88)
□(□□□)	0.5	0.60	(0.59,0.61)
□□(□□□)□□□	0.8	0.445	(0.44, 0.47)
□□(□□□□)□(□□□)□□□	1	1.00	(0.99, 1.01)

Again, the estimates are close to the truth, with the exception of □□(□□□)□□□, so the ancestral population size involving the population with the lower sample size. We conclude that the overall tree is not affected from including a population with a much lower sample size, but that population size estimates in internal branches of the tree can be affected by lower sample sizes.

Testing robustness under admixture

We also tested how admixture affected parameter estimates. We simulated three populations under a model shown in Supplementary Figure 7a. We simulated 20 chromosomes of this model under a variety of admixture rates □ , using the command line:

```
scrm 600 1 -p 12 -t 100000 -r 80000 100000000 -I 3 200
200 200 -eps 0.00125 2 3 (1-<m>) -ej 0.001875 2 1 -ej
0.00375 3 1 -seed 1
```

We then used "rarecoal mcmc", starting with the true split times and population sizes parameters of the model to estimate parameters for each simulated data set. The results are shown in Supplementary Figure 7b. Under zero admixture, the estimated parameters are very close to the true parameters, but with increasing rates of admixture, some estimates get worse, as expected, since rarecoal does not currently implement admixture. In particular, the population size of the recipient population of the admixture event (P1) is overestimated, and the older split time (t02) is underestimated. The former effect could be causing the high ancestral population size of the ancestral Spanish/Italian population (see below).

Learning the European population tree

In the following, we use three letter abbreviations for the populations studied here, which are

- FIN: Finnish from 1000 Genomes [20]

- GBR: British from 1000 Genomes
- IBS: Spanish from 1000 Genomes
- TSI: Italian from 1000 Genomes
- NED: Dutch from the GoNL data set [27]
- DMK: Danish from the GenomeDK project [28]

We started with three populations (FIN, IBS, NED) and tested all three possible tree topologies for these populations, with one global population size. The best tree, obtained via "rarecoal maxl" is ((FIN, NED), IBS) with scaled split times 0.0039 and 0.006, and a global population size of 2.3.

We then added the Danish branch and tested every possible point in the tree to join. The maximum likelihood point to join, obtained via "rarecoal find" was the Dutch branch at time 0.0028, resulting in the topology ((FIN, (NED, DMK)), IBS). We then maximized split times and a global population size on that tree using "rarecoal maxl" and found split times 0.003, 0.0038 and 0.006 with a global population size of 2.34.

Next, we added the TSI as additional population to the tree and first again checked every possible point in the tree to merge. We found that the maximum likelihood point in the tree was - surprisingly - on the Danish branch at an extremely recent time 0.0001. The second highest hit was a merge onto the Spanish branch at time 0.0023. We note that the TSI/DMK branch point may not reflect the optimal tree topology, because the branch-point search is not searching through the full space of models including individual population sizes in each branch, as is MCMC. Instead of performing MCMC on this candidate topology (TSI branching onto the DMK branch), we immediately tried the second highest merge-point with the TSI/IBS merge-point, resulting a topology ((FIN, (NED, DMK)), (IBS, TSI)). Using this candidate topology and the previous parameters as initial parameters, we then again estimated maximum likelihood parameters for this five-population tree and found parameters summarized in the following table:

Parameter	Estimate
□□□□□□□ □□	0.0024
□□□□□□□□□□□□□□ □□□	0.0032
□□□□□□□□□□	0.0049
□□□□□□□□□□□□□□ □□□□□□□□□□□□□□	0.0062
□ □□□□□□	3.15

We then allowed for separate population sizes within each branch of the tree and inferred parameters using maximization and subsequent MCMC. The results for the median estimates after MCMC are:

Parameter	Estimate
□□□□□□□ □□	0.0039
□□□□□□□□□□□□□□ □□□	0.004

[illegible]	0.0054
[illegible]	0.0064
[illegible]	0.53
[illegible]	8.23
[illegible]	6.89
[illegible]	8.37
[illegible]	1.87
[illegible]	1.05
[illegible]	0.94
[illegible]	983.25
[illegible]	2.00

Finally, we added the British population branch, by first again trying every possible point for it to merge into the tree. We found that the most likely point to merge was on the Netherland branch at time 0.0007. We used this as a starting point for another round of parameter estimation, and found that the resulting tree had two suspiciously close population splits, with a star-like split of GBR, NED and FIN. We therefore changed the topology and tried whether merging the GBR population into the ancestral (FIN, (NED, DMK))-branch would give a higher likelihood. Indeed this was the case, so the best fitting tree topology is (((FIN, (NED, DMK)),GBR),(TSI, IBS)). The final parameter estimates are:

Parameter	Estimate	95% CI
[illegible]	0.00413	(0.00412, 0.00415)
[illegible]	0.00438	(0.00436, 0.00440)
[illegible]	0.00449	(0.00447, 0.00451)
[illegible]	0.00174	(0.00168, 0.00184)
[illegible]	0.00601	(0.00599, 0.00603)
[illegible]	0.60	(0.6, 0.6)
[illegible]	4.87	(4.82, 4.94)
[illegible]	3.93	(3.8, 4.12)
[illegible]	3.26	(3.16, 3.42)
[illegible]	9.96	(9.7, 10.2)
[illegible]	1.95	(1.91, 1.99)
[illegible]	0.57	(0.55, 0.60)
[illegible]	0.71	(0.67, 0.76)
[illegible]	0.64	(0.64, 0.64)
[illegible]	997	(990, 1000)
[illegible]	1.02	(1.02, 1.02)

Since all split times are well separated considering their confidence interval, we conclude that this model represents the maximum likelihood model. If the topology was suboptimal, then the maximum likelihood result would involve star-like branch-points, with split times falling within each others confidence intervals. We also tried whether the high ancestral population size of the IBS/TSI

branch was a sub-optimal local maximum, by restarting the MCMC from a lower population size and an earlier IBS/TSI split time. This resulted in similar estimates as the ones presented above, so we conclude that this tree is the maximum likelihood tree, which is shown in Supplementary Figure 8a. The extremely high Spanish/Italian ancestral population size could an artifact of population admixture, as shown in the previous section.

Substructure in the GBR samples

As we have described in Supplementary Note 4, there is a clear substructure within the GBR samples, and so we tested each population separately with the other 5 populations. The results are shown in Supplementary Figure 8b. We first used the same tree topology as inferred for the complete GBR set above and found that it fitted well for the Orkney and Cornwall clusters, but not for the Kent cluster. We then changed the tree topology such that the Kent population was allowed to merge into the Dutch branch before other splits and obtained a significantly better fit. This suggests that the Kent population in the South of England is significantly closer to the Dutch population than both the Cornwall and Orkney group, consistent with Anglo-Saxon immigrations. This result also confirms that the second cluster in the GBR are the Kent samples, and the third cluster are the Cornish samples (see Supplementary Note 4).

Mapping individuals onto the tree

For mapping the ancient individuals onto the tree, we first generate data sets consisting of all the European individuals that went into learning the European tree, plus one additional individual. We then use the program "rarecoal find" to compute the likelihood for all branch points of the additional branch onto the tree. We vary the merge point of that additional population, over all leaf- and internal branches of the European tree, with a discretized time interval of scaled time 0.0001. In "rarecoal find", we set the options "--conditionOn" and "—minAf" to restrict the likelihood computation on sites at which the additional sample has a derived allele, and in which at least one other individual in the Reference data set has the derived allele.

We tested this approach with individuals from the 1000 Genomes project [20], which for this analysis were taken out of the reference set of FIN, GBR, IBS and TSI samples. As seen in Supplementary Figure 11, all the FIN, IBS and TSI samples fall expectedly onto the tip of their respective population branch. For the GBR individuals from Cornwall, we find that they map onto the branch of the Cornish population, as expected. When mapping individuals from Kent and Orkney, we find that they fall onto the common ancestor of all Northern European populations, similarly as the Iron Age samples.

When we use the European tree with the Kent population as British population branch, the mapping of modern samples looked different (Supplementary Figure 12). While for the FIN, IBS and TSI samples, mapping still works as expected, GBR samples from Kent do not fall onto the Kent branch. Also, one sample from Cornwall maps to the Spanish branch. The most likely explanation is that the Kent population is an admixed population and hence poorly modeled by a tree without gene flow or admixture. While the maximum-likelihood tree still places

the Kent branch closest to the Dutch population, individuals from Kent are of admixed European ancestry and hence map most likely into the ancestral branch of Northern European populations. The fact that one of the two Cornish samples maps onto the Spanish branch suggests that some Cornwall samples are genetically closer to Southern Europe than to Kent, again reflecting a more complex European history than can be modeled using simple trees.

In conclusion, we find that our approach of mapping individuals into the European tree works well for a tree with the Cornish population as British population branch, which are a relatively defined group in contrast to the samples from Kent, which have little private allele sharing (see Supplementary Figure 4a) and a large population size. In addition, it may be too admixed to be put into a simple tree phylogeny.

Supplementary Note 6 - Rarecoal Theory

The rarecoal coalescent framework

Rarecoal is a coalescent framework for rare alleles. We define rare alleles roughly by requiring i) the allele count of the derived mutation to be small, typically not larger than 10, and ii) the total number of samples to be much larger, say 100 or more. The idea is to provide a general approach of computing the joint allele frequency spectrum for rare alleles under an arbitrary demographic model under population splits and population size changes. Migration and admixture will be incorporated in the future.

Definitions

In the following, we compute the probability to observe a pattern of rare alleles seen across multiple populations, given a demographic model. In the simplest case, a demographic model is tree-like and consists of population split times and constant population sizes in each branch of the tree. Time is counted backwards in time, with $t = 0$ denoting the present and $t > 0$ denoting scaled time in the past. We denote the scaled coalescence rate (scaled inverse population size) in population k at time t by $\lambda_k(t) = N_0/N_k(t)$, where $N_k(t)$ is the population size in population k at time t, and N_0 is a scaling constant which we set to $N_0 = 20000$ for modeling human evolution.

We consider a number of P subpopulations. We define a vector $n = \{n_k\}$ for $k = 1 \ldots P$ summarizing the number of sampled haplotypes in each population. We also define vector $m = \{m_k\}$ as the set of derived allele counts at a single site in each population. As an example, consider 5 populations with 200 haplotypes sampled in each population, and a rare allele with total allele count 3, with one derived allele seen in population 2 and 2 derived alleles seen in population 3. Then we have $n = \{200, 200, 200, 200, 200\}$ and $m = \{0, 1, 2, 0, 0\}$.

Looking back in time, lineages coalesce and migrate, so the numbers of ancestral and derived alleles in the past decrease over time. In theory one needs to consider a very large state space of configurations for this process, with one state for each possible number of ancestral and derived lineages in each population. Here we make a major simplification: While we will consider the full probability distribution over the derived lineages, we will consider only the expected number of ancestral alleles over time. Specifically, we define the expected number of ancestral alleles in population k at time t as $a(t) = \{a_k(t)\}$. For the derived alleles, we define a state $x = \{x_k\}$ as a configuration of derived lineages in each population. The probability for state x at time t is defined by $b(x, t)$.

Coalescence

We now consider the evolution of the two variables $a(t)$ and $b(x, t)$ through time under the standard coalescent. We first introduce a time discretization. We define time points $t_0 = 0, \ldots t_T$. Here, $t_T = t_{max}$ should be far enough in the past to make sure that most lineages have coalesced by then with a high probability. We choose a time patterning that is linear in the beginning and crosses over to an exponentially increasing interval width. Specifically, the patterning follows this equation, inspired by the time discretization in (Li and Durbin, 2011):

$$t_i = \alpha \exp\left(\frac{i}{T} \log\left(1 + \frac{t_{max}}{\alpha}\right)\right) - \alpha. \quad (1)$$

Here, T is the number of time intervals, and α is a parameter that controls the crossover from linear to exponential scale. In practice, we use $\alpha = 0.01$, $t_{max} = 20$ and $T = 3044$, which are chosen such that

the initial step width equals one generation (in scaled units with $N_0 = 20000$), and the crossover scale is 400 generations.

Given the number of sampled haplotypes in each population n_k, and the observed number of derived alleles m_k in each population, we initialize our variables as follows:

$$a_k(t = 0) = n_k - m_k. \tag{2}$$

for each population k, and

$$b(x, t = 0) = 1 \text{ if } x_k = m_k \text{ for all } k = 1 \ldots P \tag{3}$$

$$b(x, t = 0) = 0 \text{ otherwise} \tag{4}$$

Under a linear approximation, we can compute the value of a at a time point $t + \Delta t$, given the value at time t:

$$a_k(t + \Delta t) = a_k(t) \left[1 - \frac{1}{2}(a_k(t) - 1)\lambda_k(t)\Delta t\right]. \tag{5}$$

The factor 1/2 corrects overcounting: any one coalescence takes one of two lineages out, so it should be counted half per participating lineage. We can improve this update equation slightly beyond the linear approximation: In the limit of $\Delta t \to 0$, equation 5 forms a differential equation which can be solved for finite intervals Δt:

$$a_k(t + \Delta t) = \frac{1}{1 + \left[\frac{1}{a_k(t)} - 1\right] \exp\left[-\frac{1}{2}\lambda_k(t)\Delta t\right]}. \tag{6}$$

For the derived alleles, we need to update the full probability distribution b(x, t):

$$\begin{aligned} b(x, t + \Delta t) = b(x, t) \exp\left[-\sum_k \left[\frac{x_k}{2}\lambda_k(t) + x_k a_k(t)\lambda_k(t)\right]\Delta t\right] \\ + \sum_l b(x_1 \ldots (x_l + 1) \ldots x_P, t)\left[1 - \exp\left[\frac{x_l + 1}{2}\lambda_l(t)\Delta t\right]\right] \end{aligned} \tag{7}$$

where the first term accounts for the reduction of the probability over time due to derived lineages coalescing among themselves or coalescing with an ancestral lineage, and the second term accounts for the increase from those two processes occurring in states with a higher number of derived lineages. In contrast to the equation for a(t), we cannot solve this as a differential equation and will only use this linear approximation in Δt.

Population Splits

We now consider the case where a single ancestral population splits into two separate groups at some point in time. When modelling this in a coalescent framework, we have to look at this backward in time, and thus a population split is viewed as two separate populations that join into one ancestral population at some point in time. We consider a population join backward in time from population l into population k. For the non-derived lineages, this means that after the join, population k contains the sum of lineages from population k and l:

$$a'_k(t) = a_k(t) + a_l(t) \tag{8}$$

$$a'_l(t) = 0 \tag{9}$$

where the primed variable marks the variable after the event, which will then be used as the basis for the next coalescence update.

For the derived lineages, we need to sum probabilities in the correct way. We first define a transition function that changes a state before the join to new states after the join:

$$x' = J(x), \tag{10}$$

where

$$J((\ldots x_k \ldots x_l \ldots)) = (\ldots (x_k + x_l) \ldots 0 \ldots) \tag{11}$$

We can then define the join itself as a sum over all states before the join that give rise to the same state after the join:

$$b^{\square}(x^{\square}, t) = \sum_{x, J(x) = x^{\square}} b(x, t) \tag{12}$$

The likelihood of a configuration of rare alleles

Eventually we want to compute the probability for a given configuration (n, m) observed in the present. This probability is equal to the probability that a) all derived lineages coalesce before any of them coalesces to any ancestral-allele lineage, and b) that a mutation occurred on the single lineage ancestral to all derived lineages.

We define a singleton state s^k to be the state in which only $x_k = 1$ and $x_l = 0$ for $l \neq k$. We accumulate the total probability for a single derived lineage:

$$d(t + \Delta t) = d(t) + \sum_k b(s^k)\Delta t. \tag{13}$$

Then the likelihood of the configuration under the model is

$$L(n, m) = \mu d(t_{max}) \prod_{k=1}^{P} \binom{n_k}{m_k}, \tag{14}$$

which is the total probability of a mutation occurring on a single derived lineage, times the number of ways that m derived alleles can be drawn from a pool of n samples. Note that $d(t_{max})$ depends on n, m and the demographic parameters, which we have omitted for brevity so far.

Parameter estimation

The above framework presents a way to efficiently compute the probability of observing a distribution of rare alleles, m for a large number of samples n in multiple subpopulations, given a demographic model. We can summarize the full data as a histogram of rare allele configurations. We denote the ith allele configuration by m_i and the number of times that this configuration is seen in the data by $N(m_i)$. We then write

$$L(\{N(m_i)\}|\Theta) = \prod_i L(m_i|\Theta)^{N(m_i)}, \tag{15}$$

where we have introduced a meta-parameter Θ that summarizes the entire model specification (population split times and branch population sizes), and we have made the dependency of L (eq. 14) on Θ explicit. For brevity we have omitted the sample sizes n. For numerical purpose, we always consider the logarithm of this:

$$\log L(\{N(m_i)\}|\Theta) = \sum_i N(m_i) \log L(m_i|\Theta). \tag{16}$$

The sum in equation 16 comprises all possible configurations in the genome, in principle. In practice, we only explicitly compute it for configurations between allele count 1 and 4, and replace the rest of the counts with a bulk probability:

$$\log L(\{N(m_i)\}|\Theta) = \sum_i I(AC(i))N(m_i) \log L(m_i|\Theta) + N_{other} \log L_{other}(\Theta), \tag{17}$$

where the indicator function $I(AC(i))$ gives 0 if the allele count is between 1 and 4, and 0 otherwise. The bulk count N_{other} simply counts up sites with either no variant or variants with allele count larger than 4. The bulk probability is simply:

$$L_{other}(\Theta) = 1 - \prod_i (1 - I(AC(i))L(m_i|\Theta), \tag{18}$$

With a given population tree and a given histogram of allele configuration counts $N(m_i)$, we implemented numerical optimizations over the parameters Θ to find the maximum likelihood parameters, and MCMC to estimate the posterior distributions for all parameters given the data. We usually first search for the maximum with the optimization method, which is much faster than MCMC, and then use MCMC to explore the distribution around that maximum.

Implementation

We implemented this method in the Haskell programming language as a program called "rarecoal", available from github at `https://github.com/stschiff/rarecoal`.

Supplementary References

1 Clarke, A. C., Spoerry, P. & Leith, S. Cambridgeshire: Part II Excavations at Hinxton Hall and the Genome Campus, 1993-2011: Anglo-Saxon to Medieval Settlement. (Forthcoming).

2 Lyons, A. Cambridgeshire Part I: Excavations at the Genome Campus 1993-2014: Ritual and Farming in the Cam Valley. (Forthcoming).

3 Bayliss, A., Hines, J., Hoilund-Nielsen, K., McCormac, G. & Scull, C. . (Maney Publishing, 2013).

4 Sayer, D. & Dickinson, S. D. Reconsidering obstetric death and female fertility in Anglo-Saxon England. **45**, 285-297 (2013).

5 Lucy, S. . (Sutton Pub Limited, 2000).

6 Stoodley, N. . Vol. 208 (British Archaeological Reports, 1999).

7 Sayer, D. Death and the family Developing generational chronologies. **10**, 59-91 (2010).

8 Williams, H. M. & Sayer, D. Halls of mirrors: death & identity in medieval archaeology. (2009).

9 Härke, H. Anglo-Saxon immigration and ethnogenesis. **55**, 1-28 (2011).

10 Meaney, A. L. . (London: Allen & Unwin, 1964).

11 Taylor, A., Duhig, C. & Hines, J. in 57-90.

12 Mortimer, R., Sayer, D. & Wiseman, R. in (ed S. Semple) (In Press).

13 Sayer, D. 'Sons of athelings given to the earth': 1 Infant Mortality within Anglo-Saxon Mortuary Geography. **58**, 78-103 (2014).

14 van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. **30**, E386-394, doi:10.1002/humu.20921 (2009).

15 Zheng, H. X., Yan, S., Qin, Z. D. & Jin, L. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. **2**, 745, doi:10.1038/srep00745 (2012).

16 Behar, D. M. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. **90**, 675-684, doi:10.1016/j.ajhg.2012.03.002 (2012).

17 Brandt, G. Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. **342**, 257-261, doi:10.1126/science.1241844 (2013).

18 Wei, W. A calibrated human Y-chromosomal phylogeny based on resequencing. **23**, 388-395, doi:10.1101/gr.143198.112 (2013).

19 Balaresque, P. A predominantly neolithic origin for European paternal lineages. **8**, e1000285, doi:10.1371/journal.pbio.1000285 (2010).
20 1000 Genomes Project Consortium A global reference for human genetic variation. **526**, 68-74, doi:10.1038/nature15393 (2015).
21 Skoglund, P. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. **344**, 747-750, doi:10.1126/science.1253448 (2014).
22 Jun, G. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. **91**, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).
23 Haak, W. Massive migration from the steppe was a source for Indo-European languages in Europe. **522**, 207-211, doi:10.1038/nature14317 (2015).
24 Lazaridis, I. Ancient human genomes suggest three ancestral populations for present-day Europeans. **513**, 409-413, doi:10.1038/nature13673 (2014).
25 Leslie, S. The fine-scale genetic structure of the British population. **519**, 309-314, doi:10.1038/nature14230 (2015).
26 Staab, P. R., Zhu, S., Metzler, D. & Lunter, G. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. **31**, 1680-1682, doi:10.1093/bioinformatics/btu861 (2015).
27 Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. **46**, 818-825, doi:10.1038/ng.3021 (2014).
28 Besenbacher, S. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. **6**, 5969, doi:10.1038/ncomms6969 (2015).
29 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. **475**, 493-496, doi:10.1038/nature10231 (2011).