Supplementary Text

Table of Contents

1. Linton and Hinxton Excavation Summary

Rachel Clarke, Louise Loe, Alice Lyons

Please see Supplementary Table S3 for relating Skeleton IDs with the sample names used in the paper.

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 *c.*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): 360-50 cal BC (SUERC-14246, 2155±35BP) MIDDLE-LATE IRON AGE

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: 690-900 cal AD (SUERC-20250; 1205±30BP). MIDDLE SAXON

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. 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. 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 (*c*. 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

Hinxton SK 1964 IA
Hinxton SK 241 IA-R
Hinxton SK 5518 AS
Hinxton SK 1231 AS
Hinxton SK 355 AS

Skeleton 1964: 160-26 cal BC (OxA-29573; 2039 ±27) (95.4%) IRON AGE

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.

Skeleton 1231: 170 cal BC-80 cal AD (Wk-12599; 2029±49BP) (95.4%) IRON AGE-EARLY ROMAN

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.

Skeleton 241: 666-770 cal AD (OxA-29574; 1288 ±25) (95.4%) MIDDLE SAXON

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.

Skeleton 5518: 631-776 cal AD (OxA-X-2565-12; 1320± 45) (95.4%) MIDDLE SAXON

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.

Skeleton 355: 690-881 cal AD (OxA-29572; 1230 ±25) (95.4%) MIDDLE-LATE SAXON

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.

Radiocarbon dates

Sk **318** : 1740 - 1430 cal BC (Wk-12598; 3303±68BP) (95.4%) (MBA)

Sk **1231**: 170 cal BC - 80 cal AD (Wk-12599; 2029±49BP) (95.4%) (IA-ER)

Atmospheric data from Stuiver et al. (1998); OxCal v3.5 Bronk Ramsey (2000); cub r:4 sd:12 prob usp[chron]

Sk **5518** : 631-776 cal AD (OxA-X-2565-12; 1320± 45) (95.4%)

Sk **355**: 690-881 cal AD (OxA-29572; 1230 ±25) (95.4%)

Sk **1964**: 160-26 cal BC (OxA-29573; 2039 ±27) (95.4%)

Sk **241**: 666-770 cal AD (OxA-29574; 1288 ±25) (95.4%)

Oxcal computer program (v4.2) of C. Bronk Ramsey, using the 'IntCal13' dataset

References/forthcoming publications

Lyons, A. forthcoming Hinxton, Cambridgeshire: Part I. Excavations at the Genome Campus 1993-2014: Ritual & Farming in the Cam Valley. East Anglian Archaeology Monograph.

Clarke, R. with Spoerry, P. and Leith, S. forthcoming Hinxton, Cambridgeshire: Part II Excavations at Hinxton Hall and the Genome Campus, 1993-2011: Anglo-Saxon to Medieval Settlement. East Anglian Archaeology Monograph

2. Description of Oakington Site and samples

Duncan Sayer

Please see Supplementary Table S3 for relating Skeleton IDs with the sample names used in the paper.

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 (Hines & Bayliss 2013). 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 (Dickinson 2013:228). These cemeteries are predominantly found in the south and east of England from Dorset to Northumberland with regional variation evident within the burial rite (Lucy 2000). 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 (Stoodley 1999, 2000). 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 (Sayer 2010; Williams & Sayer 2009; Härke 2011). 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 *Hochinton* and *Hochintone* in the Domesday Book of AD 1086 (VCH 1989:192-195). Oakington early Anglo-Saxon cemetery was first identified in 1926 when three burials were found as a result of cultivation (Meaney 1964). The site 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 (Taylor *et al.* 1997). 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 (Mortimer, Sayer and Wiseman, in press).

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 (Sayer 2014). 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 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 (Taylor et al. 1997:59). 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 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 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 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 (Sayer and Dickinson 2013).

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.

C14 Dating

All four samples were Carbon dated, with the following results:

Export	Pictures	BETA	Received	Due	Submitter No.	Service	Material Pretreatment	Measured Age	13C/12C	Conventional Age	2 Sigma Calibration (Click Link to Retrieve Plot)	Report Completed	QA Report	DL
		397734	Monday, December 01, 2014	Friday, December 19, 2014	OAKQUW12 1633 Vault93	AMS-Standard delivery	(bone collagen): collagen extraction: with alkali	1520 +/- 30 BP	-20.5 o/oo	1590 +/- 30 BP	Cal AD 400 to 545 (Cal BP 1550 to 1405)	Friday, December 19, 2014	26540	
		397733	Monday, December 01, 2014	Friday, December 19, 2014	OAKQUW12 1882 GR96	AMS-Standard delivery	(bone collagen): collagen extraction: with alkali	1530 +/- 30 BP	-20.5 o/oo	1600 +/- 30 BP	Cal AD 395 to 540 (Cal BP 1555 to 1410)	Friday, December 19, 2014	26540	
		397732	Monday, December 01, 2014	Friday, December 19, 2014	OAKQUW12 1870 GR95	AMS-Standard delivery	(bone collagen): collagen extraction: with alkali	1530 +/- 30 BP	-19.8 o/oo	1620 +/- 30 BP	Cal AD 385 to 475 (Cal BP 1565 to 1475) and Cal AD 485 to 535 (Cal BP 1465 to 1415)	Friday, December 19, 2014	26540	
		397731	Monday, December 01, 2014	Friday, December 19, 2014	OAKQUW12 1779 GR82	AMS-Standard delivery	(bone collagen): collagen extraction: with alkali	1490 +/- 30 BP	-20.5 o/oo	1560 +/- 30 BP	Cal AD 420 to 570 (Cal BP 1530 to 1380)	Friday, December 19, 2014	26540	

References

Dickinson, T. (2011) Overview: Mortuary Ritual. In H. Hamerow, D. Hinton, D. and S. Crawford (eds) *The Oxford Handbook of Anglo-Saxon Archaeology*. Oxford, Oxford University Press: 221-237.

Härke, H. (2011) Anglo-Saxon Immigration and Ethnogenesis. *Medieval Archaeology* 55:1-28.

Hines, J. and Bayliss, A. (eds) (2013) Anglo-Saxon Graves and Grave Goods of the 6th and 7th centuries AD: a chronological framework. York, Society of Medieval Archaeology monograph 33.

Lucy, S. (2000) *The Anglo-Saxon Way of Death*. Stroud, Sutton.

Meaney, A. (1964) A Gazetteer of early Anglo-Saxon Burial Sites. George Allen & Unwin.

Mortimer, R, Sayer, D. & Wiseman R. (In Press) Anglo-Saxon Oakington: A Central Place on the Edge of the Cambridgeshire Fen. S. Semple (ed.), *Life on the edge: Social, Political and Religious Frontiers in Early Medieval Europe*. Neue Studien zur Sachsenforschung 5 (Durham UK)

Sayer, D. & Dickinson, S.D. (2013) Reconsidering Obstetric Death and Female Fertility in Anglo-Saxon England *World Archaeology* 45(2): 285-297

Sayer, D. (2010) Death and the family: developing a generational chronology. *Journal of Social Archaeology* 10(1): 59-91.

Sayer, D. (2014). 'Sons of athelings given to the earth': infant mortality within Anglo-Saxon mortuary geography. *Medieval Archaeology* 58: 83-109

Stoodley, N. (1999) The Spindle and the Spear: A Critical Enquiry into the Construction and Meaning of Gender in the Early Anglo-Saxon Burial Rite. Oxford, Archaeopress (BAR British Series 288).

Stoodley, N. (2000) From the Cradle to the Grave: Age Organisation and the Early Anglo-Saxon Burial Rite. *World Archaeology,* 31 (3): 456-72.

Taylor, A., Duhig, C., & Hines, J. (1997). 'An Anglo-Saxon cemetery at Oakington, Cambridgeshire.' *Proceedings of the Cambridge Antiquarian Society* 86: 57–90.

Victoria County History (VCH) 1989 A History of the County of Cambridge and the Isle of Ely: Volume 9: Chesterton, Northstowe, and Papworth Hundreds: 199-204

Williams, H. and D. Sayer, (2009) Hall of Mirrors: Death and Identity in Medieval Archaeology. In D. Sayer, and H. Williams, (eds) *Mortuary Practice and Social Identities in the Middle Ages.* Exeter, The Exeter University Press: 1-22.

Oakington Acknowledgments

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.

3. DNA extraction and DNA libraries preparation

Wolfgang Haak and Bastien Llamas

Upon arrival at the Australian Centre for Ancient DNA (ACAD), bone and tooth samples from Hinxton, Linton and Oakington were documented and photographed, followed by UV-irradiation (260nm) for 20-30min. The sample surface was cleaned with a tissue soaked with commercial bleach (3.5%) and subsequently wiped with a second tissue soaked with isopropanol. After drying (2-3 minutes), 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 extraction was carried out in the clean-room facilities at ACAD, using an in-solution silica-based protocol[1]. Briefly, 200 mg of tooth or bone powder were demineralized in 4 mL 0.5M EDTA and 60 µL of Proteinase K at 37°C overnight, followed by an additional incubation with 60 µL of Proteinase K at 55°C for 1 h on the second day. DNA was subsequently bound to silica particles using a custom binding buffer made of 90% Qiagen QG buffer (including 5 M Guanidinium thiocyanate and 20 mM Tris HCl pH6.6), 220 mM Sodium acetate, 25 mM Sodium chloride, 1% Triton X 100 (all Sigma Aldrich) at a 1:4 ratio of lysate: binding buffer (vol:vol). DNA binding to silica for 1h at room temperature was followed by two washes of the silica pellet in 80% ethanol, and a final elution in 150-200 µL TE buffer (10mM Tris, 0.1mM EDTA, pH8.0) containing 0.05% Tween 20. DNA extracts were aliquoted and stored at -18°C until further use.

Double-stranded DNA libraries were prepared in the ACAD clean-room facilities using three protocols that differ slightly in the initial steps of DNA damage repair. For the Hinxton samples, libraries were prepared with truncated barcoded adapters using the 'standard' library preparation protocol by Meyer and Kircher[2]. This protocol used 20 µL of DNA extract and did not involve UDG treatment to remove characteristic ancient DNA damage, i.e. deamination of cytosines into uracils (Supplementary Table S3). From the same DNA extracts, we also prepared a second round of libraries with truncated barcoded adapters and using a UDG and endonuclease VIII damage repair treatment (USER enzyme mix, NEB) following the protocol by Briggs et al.[3]. For the Oakington and Linton samples, we built DNA libraries with truncated barcoded adapters from 25 µL of extract, and with partial DNA damage removal ('UDG-half') following a protocol recently developed by Rohland et al.[4]. The 'UDG-half' method removes damaged nucleotides except for the first and last positions of the fragments. The method is economical as it allows the assessment of DNA damage at the terminal nucleotides, while the interior of the molecules is free of DNA damage and allows unbiased variant calling for population genetic analysis.

These protocols evolved over the course of the study, especially 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). Alternatively, 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 7 cycles using full-length 7-mer indexed Illumina primers as described[2], followed by purification with Ampure and quantification using a TapeStation (Agilent). Libraries were pooled at equimolar ratios and sequenced for pre-screening purpose using an Illumina MiSeq (2x150 cycles) at ACAD, Adelaide, as well as in Harvard (see Methods). For deep sequencing, libraries were dried via evaporation at 70°C for 1h, and shipped to the Wellcome Trust Sanger Institute for deep sequencing.

1 Brotherton, P. *et al.* Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nat Commun* **4**, 1764, doi:10.1038/ncomms2656 (2013).

2 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**, pdb prot5448, doi:10.1101/pdb.prot5448 (2010).

3 Briggs, A. W. & Heyn, P. Preparation of next-generation sequencing libraries from damaged DNA. *Methods Mol Biol* **840**, 143-154, doi:10.1007/978-1-61779-516-9_18 (2012).

4 Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial UDG treatment for screening of ancient DNA. *Phil. Trans. R. Soc. B* (2014).

4. Sequencing and Read Processing

Stephan Schiffels

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 a 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 distributed samples as follows:

ID	Lanes
12880A (HI1)	4
12881A (HS1)	8
12883A (HS2)	8
12884A (HI2)	16
12885A (HS3)	4

In the second batch, we introduced 5 dark cycles into read 1 to avoid low complexity issues due to the clean room tags in the library preparation. We also included 5% PhiX sequences to increase the complexity of the first five basepairs 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). The distribution of lanes for the samples was:

ID	Lanes
15558A (O1)	8
15569A (O2)	4
15570A (O3)	10
15577A (O4)	6
15579A (L)	4

Raw Read processing

As a first processing step, we filtered out all read pairs that did not carry the correct clean room tags in the first five basepairs 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 50pb, which means that many reads contained the reverse complement of the clean room tag of the other read, and the Illumina adapters. As a last step, we removed the clean room tags and the adapters from both ends of the merged reads. Both merging and adapter trimming was done using a custom program 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 50bp) for each sample. We aligned those single reads with the program `bwa aln` [1] to the human reference, version GRCh37. We used the parameter `-l 1024` in case of the Oakinton and Linton samples, and no parameters

in case of the Hinxton samples. 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`. Finally, we 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.

DNA damage

Despite enzymatic damage repair, some low levels of DNA damage can still be found in the libraries. We used the program `mapdamage2` [2] 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 withouth UDG treatment [2].

We then used mapDamage to rescale the base quality values in the entire alignment according to the statistical model obtained from running on chromosome 20. The final bam file after damage rescaling was used for all further analysis. The bam files are available from the authors upon request.

References

[1] Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 17541760. http://doi.org/10.1093/bioinformatics/btp324

[2] Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics, 29(13), 16821684. http://doi.org/10.1093/bioinformatics/btt193

5. Mitochondrial DNA and Y chromosome analysis

Pirita Paajanen and Stephan Schiffels

Mitochondrial haplogroups

The haplogrouping was done by calling consensus sequences using samtools 0.1.19 using samtools mpileup -u -t DPR -r MT , and bcftools view -v snps version 1.1. 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 [7]. 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 are among the most common modern haplogroups in the UK. The haplogroup H1 (HI2, O1, O4, L) is found in 13,83% of modern 1000 genomes GBR samples, H (HS1, HS3) in 20,21 %, T (O3) 11.43 %, K1 (Hinx1, Hinx3) 1.06% while U5 (O1) 13,83%, [9].

Approximate times for haplogroups can be inferred from [3], and based on these, the age of the haplogroups of our samples are between 8501.8 years 1428.8 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
HI1	K1a1b1b	195	7320.3 ± 4471.4
HS1	H2a2b1	72, 195	4395.7 ± 2087.5
HS2	K1a4a1a2b		1807.5 ± 2276.1
HI2	H1ag1	152	4291.4 ± 2311.7
HS3	H2a2a1		3801.5 ± 2093.8
O1	U5a2a1	150	6094.1 ± 2635.9
O2	H1g1		1428.8 ± 1900.7
O3	T2a1a	7941	6425.8 ± 2165.4
O4	H1at1		2396.3 ± 2935.4
L	H1e	14110, 16362	8501.8 ± 2025.5

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 [4] 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 [8]. The Supplementary Table S1 in [8] 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 [5], 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.

Supplementary Table S4 lists the diagnostic genotype calls for HI1 and HI2. 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.

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. [2].

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 [1]

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 [6]. We assumed independence of the bases, and estimated

$$\hat{c} = \frac{N_{\text{alt}}}{N_{\text{cons}} + N_{\text{alt}}}.$$

If alternative alleles were found, a 95% confidence interval was computed using a binomial approximation

$$95\%CI = c \pm \sqrt{\frac{c(1-c)}{N}}.$$

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

$$P = \binom{N}{k} c^k (1-c)^{N-k},$$

where $k = 0$ and $N = N_{\text{cons}}$. In the cases where no diagnostic sites were found, the contamination could not be estimated.

Name	mtDNA Coverage	Informative Sites	N_{Cons}	N_{Alt}	contam. est.	95% CI
HI1	1145	4	5341	3	$5.6 \cdot 10^{-4}$	0–$1.1 \cdot 10^{-3}$
HS1	1020	4	4197	3	$7.1 \cdot 10^{-4}$	0–$1.5 \cdot 10^{-3}$
HS2	537	7	3290	89	0.027	0.021–0.033
HS2*	537	6	2673	10	$3.7 \cdot 10^{-3}$	$1.4 \cdot 10^{-3}$–$6.0 \cdot 10^{-3}$
HI2	2177	1	2473	13	$5.2 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$–$8.0 \cdot 10^{-3}$
HS3	587	6	4206	0	0	0–$7.1 \cdot 10^{-4}$
O1	642	3	9168	0	0	0– $3.3 \cdot 10^{-4}$
O2	652	0				
O3	410	6	35913	4	$1.1 \cdot 10^{-4}$	$2.2 \cdot 10^{-06}$–$2.2 \cdot 10^{-4}$
O4	255	0				
L	78	0				

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, and the contamination fraction is then $3.7 \cdot 10^{-3}$, similar to the other samples. 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.

References

[1] 1000 Genomes Project Consortium, Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. 491(7422):56–65, November 2012.

[2] Patricia Balaresque, Georgina R. Bowden, Susan M. Adams, Ho-Yee Leung, Turi E. King, Zo H. Rosser, Jane Goodwin, Jean-Paul Moisan, Christelle Richard, Ann Millward, Andrew G. Demaine, Guido Barbujani, Carlo Previder, Ian J. Wilson, Chris Tyler-Smith, and Mark A. Jobling. A Predominantly Neolithic Origin for European Paternal Lineages. *PLoS Biol*, 8(1):e1000285, January 2010.

[3] Doron M. Behar, Mannis van Oven, Saharon Rosset, Mait Metspalu, Eva-Liis Loogvli, Nuno M. Silva, Toomas Kivisild, Antonio Torroni, and Richard Villems. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *American Journal of Human Genetics*, 90(4):675–684, April 2012.

[4] Guido Brandt, Wolfgang Haak, Christina J. Adler, Christina Roth, Anna Szcsnyi-Nagy, Sarah Karimnia, Sabine Mller-Rieker, Harald Meller, Robert Ganslmeier, Susanne Friederich, Veit Dresely, Nicole Nicklisch, Joseph K. Pickrell, Frank Sirocko, David Reich, Alan Cooper, Kurt W. Alt, and The Genographic Consortium. Ancient DNA Reveals Key Stages in the Formation of Central European Mitochondrial Genetic Diversity. *Science*, 342(6155):257–261, October 2013.

[5] International Society of Genetic Genealogy (2014). Y-dna haplogroup tree 2014, version: Version: 10.26, date: 10 april 2015. http://www.isogg.org/tree/. Accessed: 2015-06-17.

[6] Pontus Skoglund, Helena Malmström, Ayça Omrak, Maanasa Raghavan, Cristina Valdiosera, Torsten Günther, Per Hall, Kristiina Tambets, Jüri Parik, Karl-Göran Sjögren, Jan Apel, Eske Willerslev, Jan Storå, Anders Götherström, and Mattias Jakobsson. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science (New York, NY)*, 344(6185):747–750, May 2014.

[7] Mannis van Oven and Manfred Kayser. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30(2):E386–E394, February 2009.

[8] Wei Wei, Qasim Ayub, Yuan Chen, Shane McCarthy, Yiping Hou, Ignazio Carbone, Yali Xue, and Chris Tyler-Smith. A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Research*, 23(2):388–395, February 2013.

[9] Hong-Xiang Zheng, Shi Yan, Zhen-Dong Qin, and Li Jin. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. *Scientific Reports*, 2, October 2012.

6. Rarecoal analysis

Stephan Schiffels and Richard Durbin

The rarecoal coalescent framework

This model describes 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 $\mathbf{n} = \{n_k\}$ for $k = 1 \dots P$ summarizing the number of sampled haplotypes in each population. We also define vector $\mathbf{m} = \{m_k\}$ 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 $\mathbf{n} = \{200, 200, 200, 200, 200\}$ and $\mathbf{m} = \{0, 1, 2, 0, 0\}$.

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 $\mathbf{a}(t) = \{a_k(t)\}$. For the derived alleles, we define a state $\mathbf{x} = \{x_k\}$ as a configuration of derived lineages in each population. The probability for state $\mathbf{x}$ at time t is defined by $b(\mathbf{x}, t)$.

Coalescence

We now consider the evolution of the two variables $a(t)$ and $b(\mathbf{x}, t)$ through time under the standard coalescent. We first introduce a time discretization. We define time points $t_0 = 0, \dots t_T$. 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 [1]:

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

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

$$b(\mathbf{x}, t=0) = 1 \text{ if } x_k = m_k \text{ for all } k = 1 \dots P \tag{3}$$
$$b(\mathbf{x}, t=0) = 0 \text{ otherwise} \tag{4}$$

Under a linear approximation, we can compute the value of $\mathbf{a}$ 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:

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

For the derived alleles, we need to update the full probability distribution $b(\mathbf{x}, t)$:

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

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

Population splits forward in time are joins backward in time. We consider a population join backward in time from population l into population k. For the ancestral 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:

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

where

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

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:

$$b'(\mathbf{x}', t) = \sum_{\mathbf{x}, J(\mathbf{x})=\mathbf{x}'} b(\mathbf{x}, t) \tag{12}$$

The likelihood of a configuration of rare alleles

Eventually we want to compute the probability for a given configuration ($\mathbf{n}$, $\mathbf{m}$) 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 $\mathbf{s}^k$ 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:

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

Then the likelihood of the configuration under the model is

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

which is the total probability of a mutation occurring on a single derived lineage, times the number of ways that $\mathbf{m}$ derived alleles can be drawn from a pool of $\mathbf{n}$ samples. Note that $d(t_{\max})$ depends on $\mathbf{n}$, $\mathbf{m}$ 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, $\mathbf{m}$ for a large number of samples $\mathbf{n}$ 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 $\mathbf{m}_i$ and the number of times that this configuration is seen in the data by $N(\mathbf{m}_i)$. We then write

$$\mathcal{L}(\{N(\mathbf{m}_i)\}|\Theta) = \prod_i L(\mathbf{m}_i|\Theta)^{N(\mathbf{m}_i)}, \tag{15}$$

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 $\mathbf{n}$. For numerical purpose, we always consider the logarithm of this:

$$\log \mathcal{L}(\{N(\mathbf{m}_i)\}|\Theta) = \sum_i N(\mathbf{m}_i) \log L(\mathbf{m}_i|\Theta). \tag{16}$$

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:

$$\log \mathcal{L}(\{N(\mathbf{m}_i)\}|\Theta) = \sum_i I(\mathrm{AC}(i))N(\mathbf{m}_i) \log L(\mathbf{m}_i|\Theta) + N_{\mathrm{other}} \log L_{\mathrm{other}}(\Theta), \tag{17}$$

where the indicator function $I(\mathrm{AC}(i))$ 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:

$$L_{\mathrm{other}}(\Theta) = 1 - \sum_i (1 - I(\mathrm{AC}(i))L(\mathbf{m}_i|\Theta), \tag{18}$$

With a given population tree and a given histogram of allele configuration counts $N(\mathbf{m}_i)$, 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.hs`.

Testing Rarecoal with simulated data

We defined a simple population-tree, as shown in Figure 3b of the paper. We used the SCRM simulator [2] 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. We first obtained maximum likelihood estimates of only the split times, and a globally fixed population size. Note: all times are scaled with $2N_0$ (not $4N_0$ as in the command line above), and all population sizes are scaled by N_0. This first round of maximization is summarized in the following table:

Parameter	True value	Initial value	Estimate
$t_{(0,1)}$	0.0025	0.001	0.00271
$t_{(2,3)}$	0.005	0.002	0.00242
$t_{((2,3),4)}$	0.0075	0.003	0.00452
$t_{(((0,1),((2,3)),4))}$	0.01	0.004	0.00592
N_{global}	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 MCMC:

Parameter	True Value	Median Estimate	95% CI
$t_{(0,1)}$	0.0025	0.00266	(0.00265, 0.00268)
$t_{(2,3)}$	0.005	0.00497	(0.00495, 0.00499)
$t_{((2,3),4)}$	0.0075	0.00814	(0.00812, 0.00816)
$t_{(((0,1),((2,3)),4))}$	0.01	0.00965	(0.00963, 0.00967)
N_0	0.1	0.1013	(0.1012, 0.1014)
N_1	2	2.30	(2.28, 2.31)
N_2	1	0.995	(0.992, 0.998)
N_3	5	4.98	(4.95, 5.01)
N_4	10	10.54	(10.51, 10.58)
$N_{(0,1)}$	1	0.9315	(0.9309, 0.9322)
$N_{(2,3)}$	0.5	0.6123	(0.6122, 0.6130)
$N_{((2,3),4)}$	0.8	0.4648	(0.4647, 0.465)
$N_{(((0,1),((2,3)),4))}$	1	0.928	(0.92, 0.934)

In most parts of the tree, the estimates are close to the truth, with one exception: the worst fit parameter is $N_{((2,3),4)}$, the ancestral population size in the branch preceeding the second split, which is about 40% too low. This may be due to the fact that this branch is relatively short and the subtree below has relatively large population sizes, which are both causes of relatively low amounts of genetic drift and consequently relatively weak information in the data about parameters.

Learning the European population tree

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 is ((FIN, NED), IBS) with scaled split times 0.0039 and 0.006, and a global populaton isze 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 was the Dutch branch at time 0.0028, resulting in the topology ((FIN, (NED,

DMK)), IBS). We then maximized split times and global population size on that tree 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. We decided this to be some artifact of the fixed population sizes and chose the second-highest merge-point, which was the Spanish branch at time 0.0023, resultin a topology ((FIN, (NED, DMK)), (IBS, TSI)). Using this merge-point 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**
$t_{\texttt{(NED,DMK)}}$	0.0024
$t_{\texttt{(FIN,(NED, DMK))}}$	0.0032
$t_{\texttt{(IBS,TSI)}}$	0.0049
$t_{\texttt{((FIN,(NED,DMK)),(IBS,TSI))}}$	0.0062
N_{global}	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 are as follows:

Parameter	**Estimate**
$t_{\texttt{(NED,DMK)}}$	0.0039
$t_{\texttt{(FIN,(NED, DMK))}}$	0.004
$t_{\texttt{(IBS,TSI)}}$	0.0054
$t_{\texttt{((FIN,(NED,DMK)),(IBS,TSI))}}$	0.0064
N_{FIN}	0.53
N_{IBS}	8.23
N_{TSI}	6.89
N_{NED}	8.37
N_{DMK}	1.87
$N_{\texttt{(NED,DMK)}}$	1.05
$N_{\texttt{(FIN,(NED, DMK))}}$	0.94
$N_{\texttt{(IBS,TSI)}}$	983.25
$N_{\texttt{((FIN,(NED,DMK)),(IBS,TSI))}}$	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 an essesntially star-like split of GBR, NED and FIN. We therefore changed the topology and tried whether merging the GBR population into the Finnish branch would give a higher likelihood. Indeed this was the case, so the best fitting tree topology is (((FIN, GBR), (NED, DMK)), (TSI, IBS)). The final parameter estimates are:

Parameter	**Median Estimate**	**95% CI**
$t_{\texttt{(NED,DMK)}}$	0.00372	(0.00370, 0.00373)
$t_{\texttt{(FIN, GBR)}}$	0.00399	(0.00398, 0.004)
$t_{\texttt{((FIN, GBR),(NED, DMK))}}$	0.00417	(0.00415, 0.00418)
$t_{\texttt{(IBS,TSI)}}$	0.00238	(0.00237, 0.00240)
$t_{\texttt{((FIN, GBR),(NED,DMK)),(IBS,TSI))}}$	0.00605	(0.00603, 0.00607)
N_{FIN}	0.54868	(0.54863, 0.54874)
N_{GBR}	4.353	(4.347, 4.358)
N_{IBS}	4.910	(4.908, 4.913)
N_{TSI}	4.3263	(4.3253, 4.3272)
N_{NED}	10.500	(10.488, 10.508)
N_{DMK}	1.755	(1.741, 1.771)
$N_{\texttt{(NED,DMK)}}$	1.060	(1.059, 1.061)
$N_{\texttt{(FIN, GBR)}}$	0.85	(0.85, 0.85)
$N_{\texttt{((FIN, GBR),(NED, DMK))}}$	0.86	(0.86, 0.86)
$N_{\texttt{(IBS,TSI)}}$	998	(992, 1000)
$N_{\texttt{((FIN, GBR),(NED,DMK)),(IBS,TSI))}}$	1.8149	(1.8137, 1.8169)

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.

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 compute the likelihood for a family of models, which are all composed of the original model learned for the European populations, plus one more population that merges onto the tree. We vary only the merge point of that additional population, over all leaf- and internal branches of the Europena tree, with a discretized time interval of scaled time 0.0001. In this likelihood computation, we deviated from the standard likelihood (equation 17) in one detail: We only explicitly fitted sites in which the ancient samples carried the derived allele. All other sites were accumulated into the bulk number (N_{other}) alongside variants with allele count higher than 4 and sites without variants.

We tested this approach with individuals from the 1000 Genomes project [3], which for this analysis were taken out of the reference set of FIN, GBR, IBS and TSI samples. As seen in Extended Data Figure 8, all of the 8 individuals shown map onto the tip of their population branch, as exptected for samples that belong to those reference populations. For the GBR individuals, we also noticed some systematic deviations, as shown in Extended Data Figure 8, with some samples mapping to the common ancestor of all Northern European populations. We believe this is due to population structure within the GBR samples in 1000 Genomes, which were sampled from three locations (Kent, Cornwall and Orkney). Because our European tree assumes one panmictic population for those subpopulations we expect some samples to not be represented by this population. Note that we do not know the sampling locations of any individual in 1000 Genomes, so cannot separate them on the European tree in the first place.

References

[1] Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. http://doi.org/10.1038/nature10231

[2] Staab, P. R., Zhu, S., Metzler, D., and Lunter, G. (2015). scrm: efficiently simulating long sequences using the approximated coalescent with recombination. Bioinformatics, btu861. http://doi.org/10.1093/bioinformatics/btu861

[3] 1000 Genomes Project. (2015). A global reference for human genetic variation. Nature (in Revision).