Open access • Posted Content • DOI:10.1101/2021.02.17.431573

Inferring population histories for ancient genomes using genome-wide genealogies — Source link

Leo Speidel, Leo Speidel, Lara M. Cassidy, Robert W. Davies ...+3 more authors

Institutions: Francis Crick Institute, University College London, Trinity College, Dublin, University of Oxford

Published on: 17 Feb 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Population and Imputation (genetics)

Inferring population histories for ancient genomes using genome-wide genealogies

Leo Speidel[1,2], Lara Cassidy[3], Robert W. Davies[4],
Garrett Hellenthal[2], Pontus Skoglund[1], Simon R. Myers[4,5]

[1]Francis Crick Institute, London, UK

[2]Genetics Institute, University College London, London, UK

[3]Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Republic of Ireland

[4]Department of Statistics, University of Oxford, Oxford, UK

[5]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

Abstract

Ancient genomes anchor genealogies in directly observed historical genetic variation and contextualise ancestral lineages with archaeological insights into their geography and cultural associations. However, the majority of ancient genomes are of lower coverage and cannot be directly built into genealogies. Here, we present a fast and scalable method, *Colate,* the first approach for inferring ancestral relationships through time between low-coverage genomes without requiring phasing or imputation. Our approach leverages sharing patterns of mutations dated using a genealogy to infer coalescence rates. For deeply sequenced ancient genomes, we additionally introduce an extension of the Relate algorithm for joint inference of genealogies incorporating such genomes. Application to 278 present-day and 430 ancient DNA samples of >0.5x mean coverage allows us to identify dynamic population structure and directional gene flow between early farmer and European hunter-gatherer groups. We further show that the previously reported, but still unexplained, increase in the TCC/TTC mutation rate, which is strongest in West Eurasia today, was already present at similar strength and widespread in the Late Glacial Period ~10k-15k years ago, but is not observed in samples >30k years old. It is strongest in Neolithic farmers, and highly correlated with recent coalescence rates between other genomes and a 10,000-year-old Anatolian hunter-gatherer. This suggests gene-flow among ancient peoples postdating the last glacial maximum as widespread and localises the driver of this mutational signal in both time and geography in that region. Our approach should be widely applicable in future for addressing other evolutionary questions, and in other species.

1 Introduction

Genetic variation is shaped through evolutionary processes acting on our genomes over hundreds of millennia, including past migrations, isolation by distance, mutation or recombination rate changes, and natural selection. Such events are reflected in the genealogical trees that relate individuals back in time. While these are unobserved, recent advances have made their reconstruction from genetic variation data increasingly feasible, with the most scalable methods now able to build trees for many thousands of individuals (Speidel et al. 2019; Kelleher et al. 2019). This has enabled powerful inferences of our genetic past (Rasmussen et al. 2014; Kelleher et al. 2019; Speidel et al. 2019).

Ancient genomes provide a direct snapshot of historical genetic variation, and so add substantial information compared to genealogies built only from modern-day samples. We introduce an extension to the *Relate* algorithm to enable the incorporation of samples of variable ages. We use this approach to reconstruct joint genealogies of the Simon's Genome Diversity Project (SGDP) dataset (Mallick et al. 2016) and 14 previously published high-coverage ancient humans covering diverse ancestries and sampled across the last 45k years (Fu et al. 2014; Lazaridis et al. 2014; Gallego-Llorente et al. 2015; Jones et al. 2015; Broushaki et al. 2016; Sikora et al. 2017; de Barros Damgaard et al. 2018; Günther et al. 2018; Sikora et al. 2019; Cassidy et al. 2020). These genealogies capture the shared population histories of present-day and ancient humans. In particular, they allow identification of inbreeding, directional migration, and estimation of coalescence rates between individuals, analysis of the age and spread of individual mutations, and in future might be used to infer natural selection (Speidel et al. 2019). A similar approach could also be applied to other species.

The joint inference of genealogies for ancients and moderns currently requires accurate diploid genotypes, and thus excludes the majority of ancient human genomes, because these have lower sequencing coverage. One central set of questions for such samples involve estimation of their joint genetic history: their historical relationships with one another through time, reflected in their varying coalescence rates through time. These coalescence rates can be estimated using a number of methods (Gutenkunst et al. 2009; Li and Durbin 2011; Schiffels and Durbin 2014; Terhorst et al. 2017; Kamm et al. 2020), as well as our updated *Relate* approach, but to date none of these have been designed to work for low-coverage genomes. We have therefore developed a fast and scalable method, *Colate*, for inferring coalescence rates between low-coverage genomes without requiring phasing or imputation. *Colate* leverages the distributions of mutational ages from a *Relate*-inferred genealogy to construct a likelihood based on the changing pattern of sharing of mutations through time, which we maximise using an Expectation-Maximisation (EM) algorithm. The method can calculate coalescence rates between any number of samples. Running Colate involves two steps: first, a preprocessing step whose complexity is linear in sample size and genome length and secondly a constant time (~5 seconds, **Methods**) analysis step to run the EM algorithm .

We applied *Colate* to 430 genomes of >0.5x coverage spanning the late Paleolithic, Mesolithic, Neolithic, and more recent epochs across many regions outside Africa (SI Table). Among other findings, we readily identify genetic clusters corresponding to hunter-gatherers (HGs), Early farmers, and the Late Neolithic-Bronze age transition in Europe, and map out the coalescence rates of modern humans worldwide with these ancient samples. We show that these indicate localised structure which converges back in time, and characterise dramatic population replacements in Ireland within the space of 3,000 years. A strength of the Relate and Colate approaches is that they extrapolate relationships of individuals to the past where data is comparatively sparse. We find evidence of directional gene flow between European HG groups across Europe predating the Neolithic, which is more widespread than previously identified.

Finally, we leverage our *Relate*-inferred genealogies and *Colate*-inferred coalescence rates to quantify the previously reported but unexplained elevation in TCC to TTC mutation rate (Harris 2015) in all SGDP individuals and 161 ancient individuals of >2x mean coverage, providing a finer-scale geographic and temporal mapping of this signal than previously available. We show that the signal shows a remarkable 96% correlation with coalescence rates with an early Anatolian farmer from the pre-pottery Neolithic (Kılınç et al. 2016). While absent in samples from >34,000 years before present (YBP), it was already widespread among HGs in Late Glacial West Eurasia, and shows no increase in strength over the last 10,000 years, suggesting that the driver of this mutational signature was extinct by the Holocene. This strong localisation of the signal in both time and space suggests either a genetic cause, or a somehow tightly focussed environmental cause. Moreover, we hypothesise that these excess TCC/TTC mutations spread via gene flow through ancestors of ancient Anatolia into HG groups across Western Eurasia before the expansion of farming, perhaps associated with a link between the Near East and Late Upper Paleolithic Europe that started with the Bølling–Allerød interstadial warming period (Fu et al. 2016).

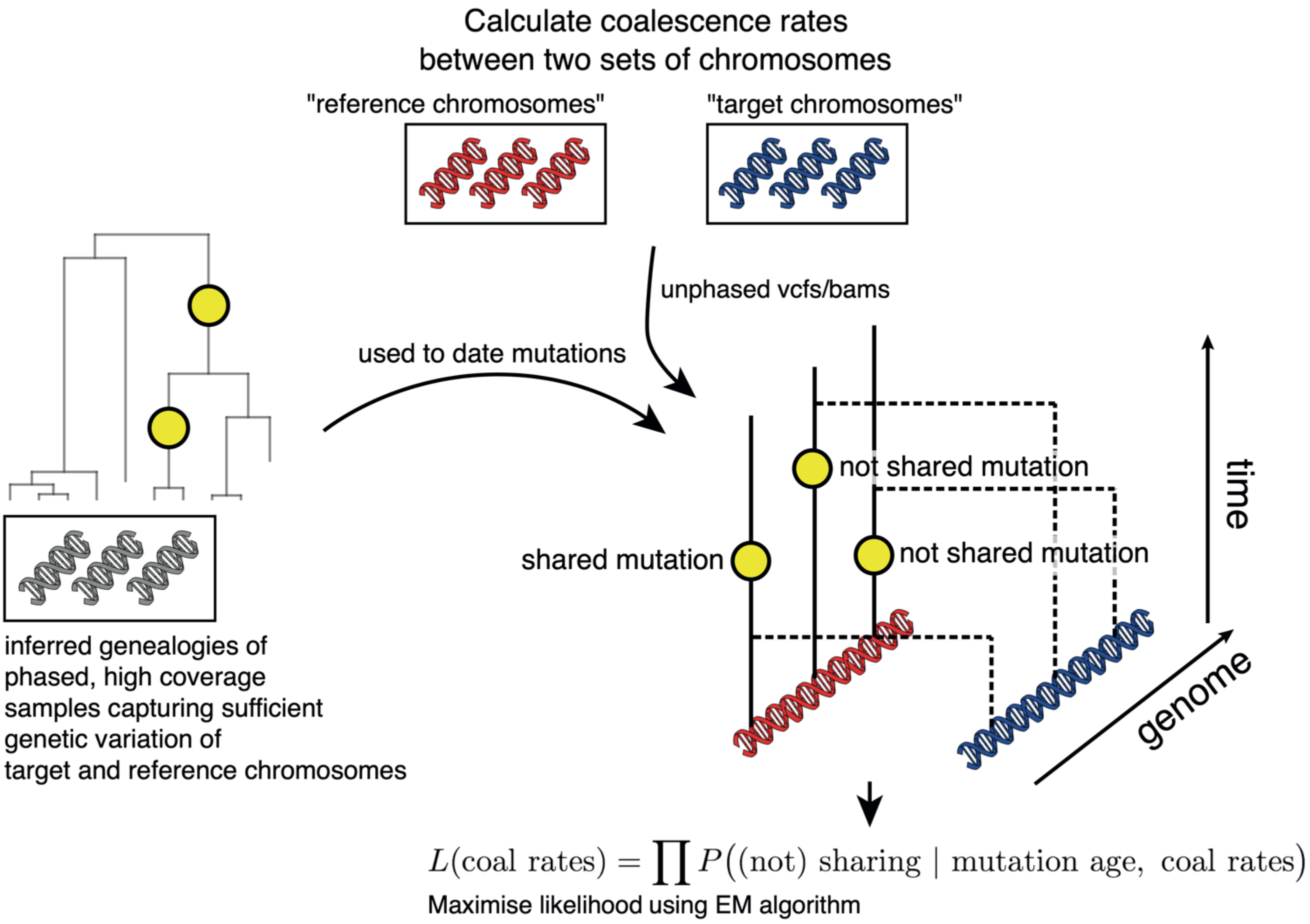
Calculate coalescence rates
between two sets of chromosomes
"reference chromosomes"
"target chromosomes"
unphased vcfs/bams
used to date mutations
not shared mutation
shared mutation
not shared mutation
time
genome
inferred genealogies of
phased, high coverage
samples capturing sufficient
genetic variation of
target and reference chromosomes
$L(\text{coal rates}) = \prod P\big((\text{not})\ \text{sharing} \mid \text{mutation age},\ \text{coal rates}\big)$
Maximise likelihood using EM algorithm

Figure 1

Colate calculates coalescence rates between two sets of chromosomes, labelled target and reference (main text). The method proceeds by recording for each mutation carried by a reference chromosome, whether it is shared in the target chromosomes. This information is summarised in a likelihood, constructed by multiplying over SNPs, such that no phase information is required. Whenever more than one chromosome is available at any given site, we multiply across chromosomes. The likelihood is maximised using an expectation-maximisation algorithm.

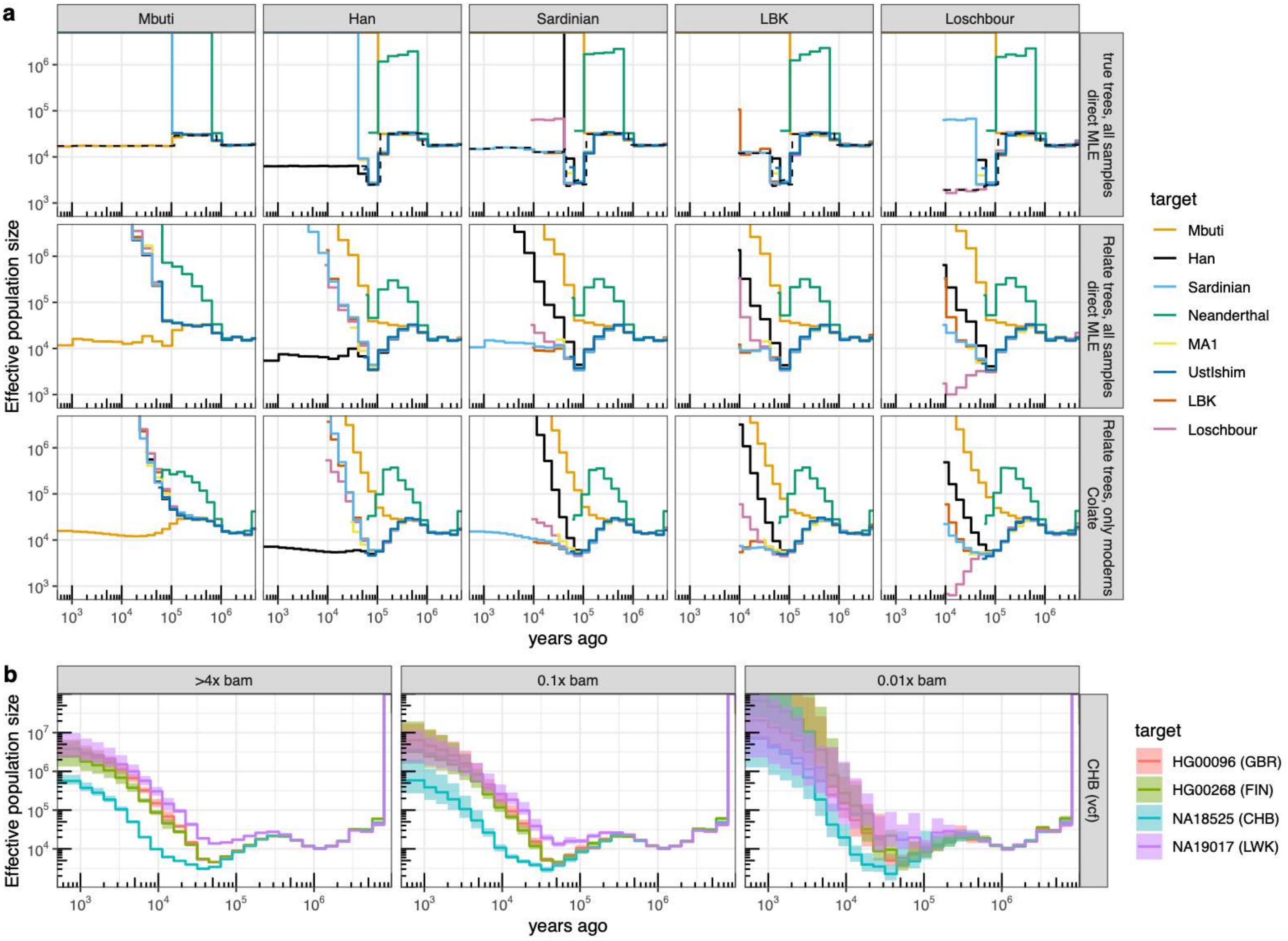

Figure 2

a, Simulation emulating real human groups, including three modern human groups (Mbuti, Han, and Sardinian) with 100 diploid sequences each, and five diploid ancient genomes. We calculated coalescence rates between groups using true genealogical trees of all samples (true trees; direct MLE), inferred *Relate* trees of all samples (*Relate* trees; direct MLE), and *Colate*, where the reference genealogy included all modern human groups but not the ancient samples. For the direct MLEs, coalescence rates are symmetric with respect to target and reference group assignment; for *Colate*, each panel corresponds to a fixed reference group, with different coloured lines showing different target groups. Five reference groups are shown here, see Supplementary Figure 3 for remaining groups. Dashed lines show true within-group population sizes. **b,** *Colate*-inferred coalescence rates between four 1000 Genomes Project samples (HG0096, HG00268, NA18525, NA19017) and the remaining 1000 Genomes samples of Han Chinese in Beijing (CHB). (see Supplementary Figure 3 for rates to YRI and CEU). The target samples are given as reference-aligned read data downsampled to 4x, 0.1x, and 0.01x mean coverage. Confidence intervals are constructed using 100 block bootstrap iterations with a block size of 20Mb.

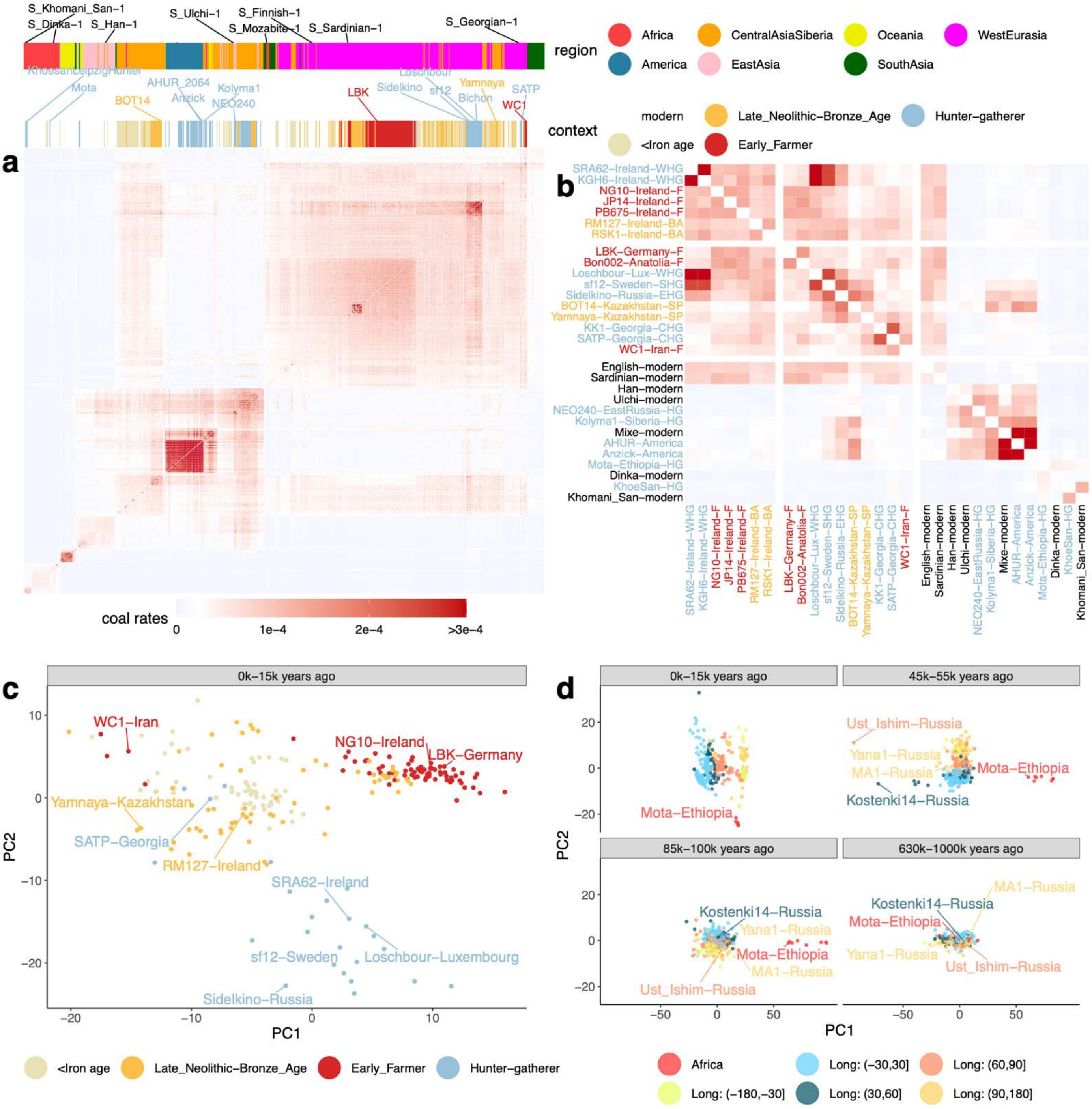

Figure 3

a, Matrix of Colate-inferred pairwise coalescence rates for all modern SGDP individuals and ancient individuals in the most recent epoch 0-15,000 years before present (YBP). **b**, Highlighted subset of samples from **a.** Sample names are coloured by context. Abbreviations in sample names are WHG: Western hunter-gatherer, SHG: Scandinavian hunter-gatherer, EHG: Eastern hunter-gatherer, CHG: Caucasus hunter-gatherer, F: farmer, BA: Bronze Age, SP: Steppe Pastoralists **c**, Principal component analysis (PCA) on pairwise coalescence rates of ancient individuals in West Eurasia in epoch 0 – 15,000 YBP, coloured by context. **d**, PCA on pairwise coalescence rates for four epochs, coloured by Longitude outside Africa. In all PCAs, we standardised columns in each matrix of coalescence rates and applied the R function prcomp to calculate PCs.

2 New approaches

2.1 Extending *Relate* to work with non-contemporary samples

We extend our previously developed method, *Relate*, for inference of genealogical trees genome-wide for large sample sizes (Speidel et al. 2019) to work with ancient genomes (Supplementary Information). A key aspect of non-contemporary samples is that, when these samples have known ages, these impose hard constraints on the times of coalescence events. Our updated tree builder restricts which lineages can coalesce by assigning a preliminary date to each coalescence event and only allows coalescences of non-contemporary samples with lineages that predate its age. Branch lengths are sampled using a Markov-Chain Monte Carlo sampler, with modified proposal distributions to allow for non-contemporary samples. As before, we sample branch lengths from a posterior distribution that fixes tree topology and combines the likelihood of observing a certain number of mutations on a branch and a coalescent prior with piecewise-constant effective population sizes through time.

2.2 Inferring coalescence rates for low-coverage genomes using *Colate*

Colate calculates coalescence rates between a set of "target" and a set of "reference" chromosomes by leveraging mutations dated using an inferred genealogy. This genealogy may (or may not) have overlapping samples with the target and reference chromosome sets (Figure 1, **Methods** and Supplementary Information). Both the target and reference chromosomes may be specified as BCF files containing genotypes, or as BAM files containing reference-aligned reads. The latter is particularly useful for low-coverage sequencing data, where accurate genotype calling is not possible. For ancient genomes, we specify a sample date. In practise, we often specify two different individuals as the target and reference, and obtain the coalescence rates between this pair, although it is also possible to group samples.

The *Colate* likelihood uses as input data whether each mutation carried by a reference chromosome is shared, or not shared, with a target chromosome. Sharing indicates that coalescence between the two chromosomes happened more recently than the age of this mutation, whereas non-sharing indicates that coalescence happened further in the past, assuming each mutation occurs only once (the infinite-sites model), and so an exact likelihood can be calculated, given coalescence rates between the individuals from whom these chromosomes are taken (**Methods**). We multiply this likelihood across sites and therefore do not require genomes to be phased; in low-coverage data, we additionally multiply across pairs of reads. This likelihood is then maximised using an expectation-maximisation

(EM) algorithm (**Methods,** Supplementary Information). Our implementation reduces computation time by using a discrete time grid to record sharing and non-sharing of mutations through time, reducing the computation time of the EM algorithm. As a result, computation time is independent of both sample size and genome lengths once the data is preprocessed, and typically on the order of 5 seconds (~40 seconds including parsing the data, Supplementary Figure 1).

We observe high accuracy of *Colate* and *Relate*-inferred coalescence rates using the stdpopsim package (Adrion et al. 2020), on simulated data following a zigzag demographic history (Supplementary Figure 2) as well as a multi-population model of ancient Eurasia, which was fitted using real human genomes (Kamm et al. 2020) (Figure 2**a**, Supplementary Figure 3) (**Methods;** see (Speidel et al. 2019) for comparison of *Relate* to other methods). We further evaluated *Colate*'s performance on low-coverage sequencing data by downsampling high-coverage genomes of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015). Although uncertainty increases as coverage decreases, *Colate* recovers meaningful coalescence rate estimates even between a sequence of 0.01x mean coverage and high-coverage sequences specified as a VCF (Figure 2**b**), or between two low coverage sequences of 0.1x mean coverage (Supplementary Figure 4).

3 Results

3.1 *Relate* and *Colate* applied to 278 SGDP moderns and 430 ancients

We inferred joint genealogies of 278 modern-day individuals of the Simons Genome Diversity Project and 14 previously published high coverage genomes of ancient individuals of >8x mean coverage, which we collectively rephase using Shapeit4 (Delaneau et al. 2019) and the 1000 Genomes Project reference panel (**Methods**). Tree topologies were constructed using all mutations except CpG dinucleotides, but branch length inference used transversions only, so as to avoid confounding due to deamination errors in the ancient genome sequences (**Methods**). Furthermore, we estimate pairwise-coalescence rates for 430 ancient individuals of >0.5x mean sequencing coverage using *Colate* (SI Table). For *Colate*, we use a *Relate*-inferred genealogy of the SGDP samples to date mutations, sampling one haplotype from each individual to remove the effects of recent inbreeding and restrict our analysis to transversions (**Methods**).

3.2 PCA on *Colate*-inferred coalescence rates captures dynamic population structure

Colate-inferred coalescence rates demonstrate intricate relationships that vary geographically and through time and manifest vast migrations and, in places, repeated population replacements (Figure 3**a,b**). In the recent past (0-15KY), populations are separated based on both geography and sample age (Figure 3**a,b**): there are extremely low coalescence rates between continental regions (excepting W. Eurasia and Central Asia, which show patterns indicating migration). Taking samples from Ireland as one example (Figure 3**b**), previous work has indicated repeated partial or complete population replacements, first of Mesolithic HGs by Neolithic farmers, and then in the Bronze age by migrants related to people from the Western steppe (Cassidy et al. 2016). Using *Colate*, the earliest Irish Mesolithic samples have highest coalescence rates with, and similar relatedness to other groups as, West European hunter-gatherers (e.g. Loschbour). Neolithic Irish samples show much lower affinity to these HGs, but are closely similar to other European farmers (e.g. LBK, an early farmer from Germany). Bronze age Irish samples again show more similarity to HGs, but now *Eastern* European HGs (and other Eastern European groups), and in this and other respects they resemble the Yamnaya, a possible source group (Figure 3**b**); however they retain some farmer-like haplotypes not present in the Yamnaya sample. Comparing across the whole dataset, we observe that Irish ancient genomes are closest to other Irish ancients from within the same time period (Supplementary Figure 5, 6). This implies that finer scale, regional stratification existed within the HGs, Neolithic farmers, and Bronze age samples, but there is no clear evidence of continuity across periods, suggesting this arose independently repeatedly. We also identify clear substructure among European HGs, consistent with previous findings (Lazaridis et al. 2014) and pairwise F2 statistics (Supplementary Figure 7); this structure corresponds to a divide of Western, Eastern, Scandinavian, and Caucasus HGs among our samples in Europe.

One approach to visualise the diverse signals in these data is to adapt the widely used PCA approach, but now using coalescence rates within particular epochs (Figure 3**c,d** show the first two PCs for selected epochs). Structure is not seen in the deep past (>630k years before present (YBP)) but in distinct epochs we observe separation first of African (e.g., Mota) and non-African individuals, and by 45-55k YBP, a separation between West and East Eurasians, as well as a stronger split with Ust'-Ishim (Fu et al. 2014), a 45k-year-old Siberian individual who also appears slightly closer to East Eurasians compared to later European samples, such as Kostenki14 (Seguin-Orlando et al. 2014) and Sunghir3 (Sikora et al. 2017), who are closer to West Eurasians. In the most recent epoch (0-15k YBP), our PCA mirrors geography globally (Novembre et al. 2008), but reflects different ancestries more strongly within smaller regions; for instance, we detect three clusters, corresponding to Mesolithic HGs, Neolithic farmers, and Bronze/Iron age individuals in Europe (Figure 3**c**). The Bronze age cluster falls closer to Steppe Pastoralists from the Pontic-Caspian Steppe (e.g., Yamnaya), consistent with previously reported gene flow from this region into Bronze age Europe (Allentoft et al. 2015; Haak et al. 2015). Overall, these inferences seem in strong agreement, across time and space, with previous specific analyses of these samples.

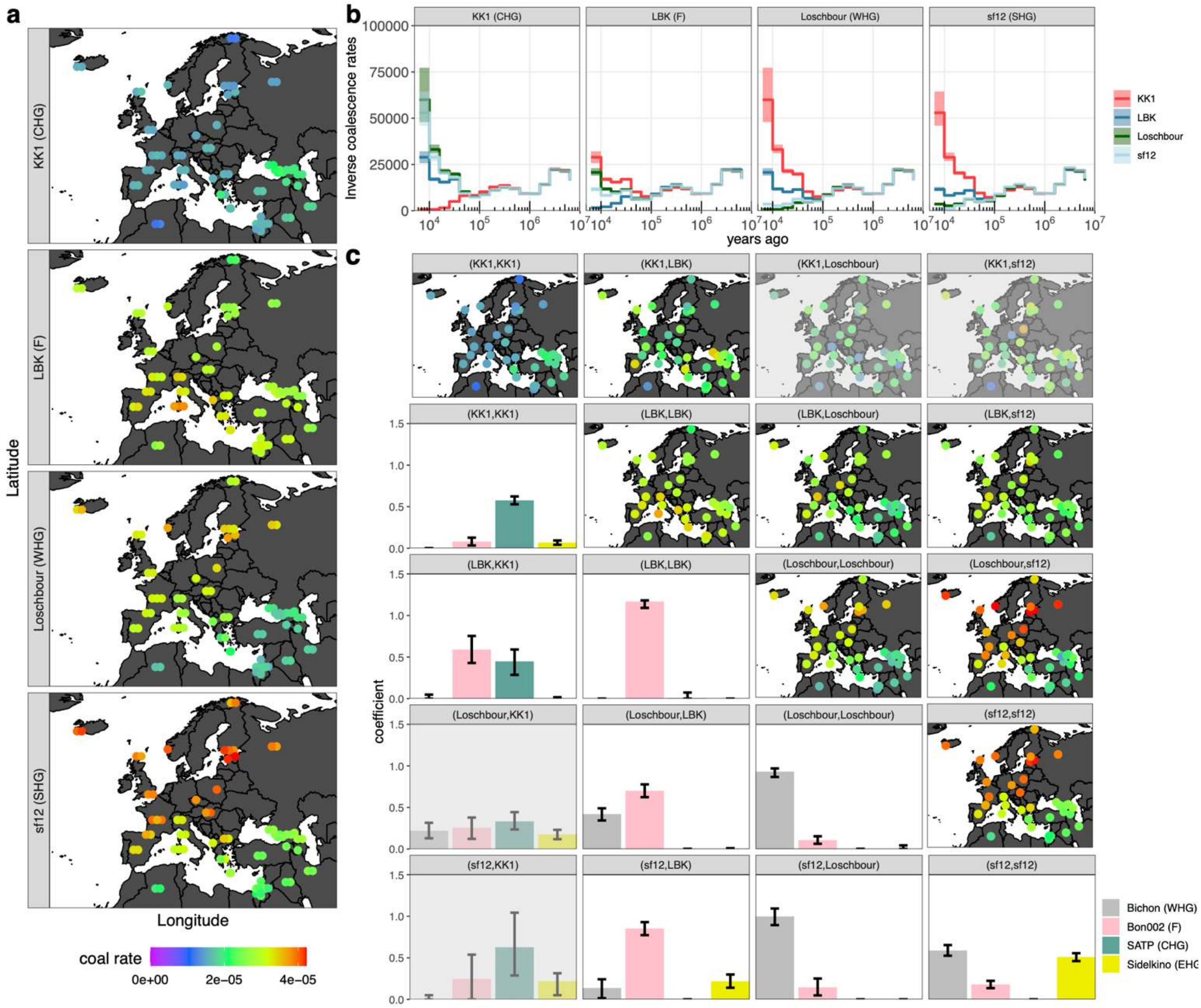

Figure 4

a, Map showing *Relate*-inferred coalescence rates of a 9700-year-old Caucasus HG (KK1), 7200-year-old early European farmer (LBK), a nearly 8000-year-old Western hunter-gatherer (Loschbour), and a 9000-year-old Scandinavian HG to SGDP modern individuals. The coalescence rates shown in the map correspond to the epoch 16k-25k YBP. b, *Relate*-inferred inverse coalescence rates (effective population sizes) for KK1, LBK, Loschbour, and sf12 to themselves and each of the other four individuals. **c**, Maps in top diagonal show *Relate*-inferred coalescence rates of lineages with descendants shown by map labels to SGDP moderns in same epoch as in **a.** Bottom diagonal shows non-linear least squares coefficients obtained by fitting coalescence rates of lineages with descendants given by map labels to SGDP moderns as a mixture of *Colate*-inferred coalescence rates of Bichon (Western HG), Bon002 (Anatolian), SATP (Caucasus HG), Sidelkino (Eastern HG) with SGDP moderns (**Methods**). Panels involving KK1 and Loschbour or sf12 are partially greyed out, as there is little recent gene-flow between these groups. Confidence intervals show 2.5 and 97.5 percentiles obtained from 1000 bootstrap samples.

3.3 Relationship of European hunter-gatherer groups to Neolithic farmers

While there is strong evidence for Anatolian farmers partially replacing HG ancestry across Europe in the Neolithic (Haak et al. 2010), the deeper relationship of ancestors of these Anatolian farmers to European HGs in the Late Upper Paleolithic is not fully understood. We therefore assess these deep relationships between early European farmers, Western, Scandinavian, and Caucasus HGs built into our Relate genealogies. These groups show distinct footprints in present-day Europeans, consistent with previous findings (Figure 4). We observe a South-North cline, with the highest farmer-like ancestry observed in Sardinians (Figure 3**b**), while Western and Scandinavian HG-like ancestry is highest in northern European groups and Caucasus HG-like ancestry is concentrated around present-day Georgia (Lazaridis et al. 2014; Skoglund et al. 2014; Jones et al. 2015).

Caucasus HGs have previously been modelled as forming a clade with early farmers that is deeply diverged from Western and Scandinavian HGs (~46k YBP) (Jones et al. 2015). Our pairwise coalescence rates among samples confirm that Western and Scandinavian HGs form a clade relative to Caucasus HGs (KK1), with a consistent split time and almost no recent coalescences observed between these groups.

However, patterns observed for early farmers (LBK) imply a non-tree-like group relationship involving migration (Figure 4**b**): Caucasus HGs show greater affinity to Neolithic farmers than to Western or Scandinavian HGs in recent epochs, but this is not reciprocated by early farmers who have higher coalescence rates to Western and Scandinavian HGs than to Caucasus HGs. Recent studies have demonstrated that the major ancestral component of Western HGs only became widespread in Northern and Western Europe after 14k YBP and harbours an increased affinity to Anatolian and Caucasus populations, relative to earlier European HGs (Fu et al. 2016), suggesting an expansion of peoples from Southeast Europe or the Near East following the Last Glacial Maximum (LGM).

To test for evidence of migration between ancestors of early farmers and other European HGs, we examine lineages that are formed recently (<50k YBP) through a coalescence of individuals from each group. If this coalescence happened more recently than the split time of groups A and B, these lineages are expected to represent migrants from one population to another. If recent migration is purely directional from group A into group B, such lineages will always come from group A in the past and will behave like a typical group A lineage back in time. To evaluate whether these recently coalesced lineages are more similar to a typical lineage ancestral to group A or group B, we compare their coalescence rates to other individuals, as this should distinguish their affiliation (group A lineages can be characterised by coalescing more rapidly with some individuals compared to group B lineages, **Methods**). To gain power, we calculate the coalescence rates to each non-African SGDP modern sample and fit these using non-negative least squares against *Colate*-inferred coalescence rates of four individuals representing independent samples from similar, but older groups: ancient Anatolia (Bon002) (Kılınç et al. 2016), Western HGs (Bichon) (Jones et al. 2015), Eastern HGs (Sidelkino) (de Barros Damgaard et al. 2018), and Caucasus HGs (SATP) (Jones et al. 2015) (Figure 4**c, Methods**). This will fit these recently coalesced lineages as a mixture of four potential surrogate source populations. We rescaled *Colate* coalescence rates according to Supplementary Figure 9 to match overall levels of coalescence rates between *Colate* and *Relate*.

Encouragingly, we find that lineages ancestral to the two haplotypes of the same individual (not indicating migration) are well captured by one respective ancestry in our regression in three cases and suggests these are reasonable surrogates. The exception is the Scandinavian HG (sf12) who we fit as an approximately equal mixture of Eastern and Western HGs, as previously reported (Günther et al. 2018). The highest recent coalescence rates across groups are between the Western and Scandinavian HG: recently coalesced lineages between these samples appear very similar to Western HGs (Figure 4**c**), indicating strong directionality of gene-flow, from Western HGs into Scandinavia. In contrast, gene-flow between Western HGs (Loschbour) and early farmers (LBK) appears strongly bidirectional in our analysis, as do lineages ancestral to LBK and Scandinavian or Caucasus HGs, therefore suggesting widespread migration between ancestors of these groups predating the European Neolithic.

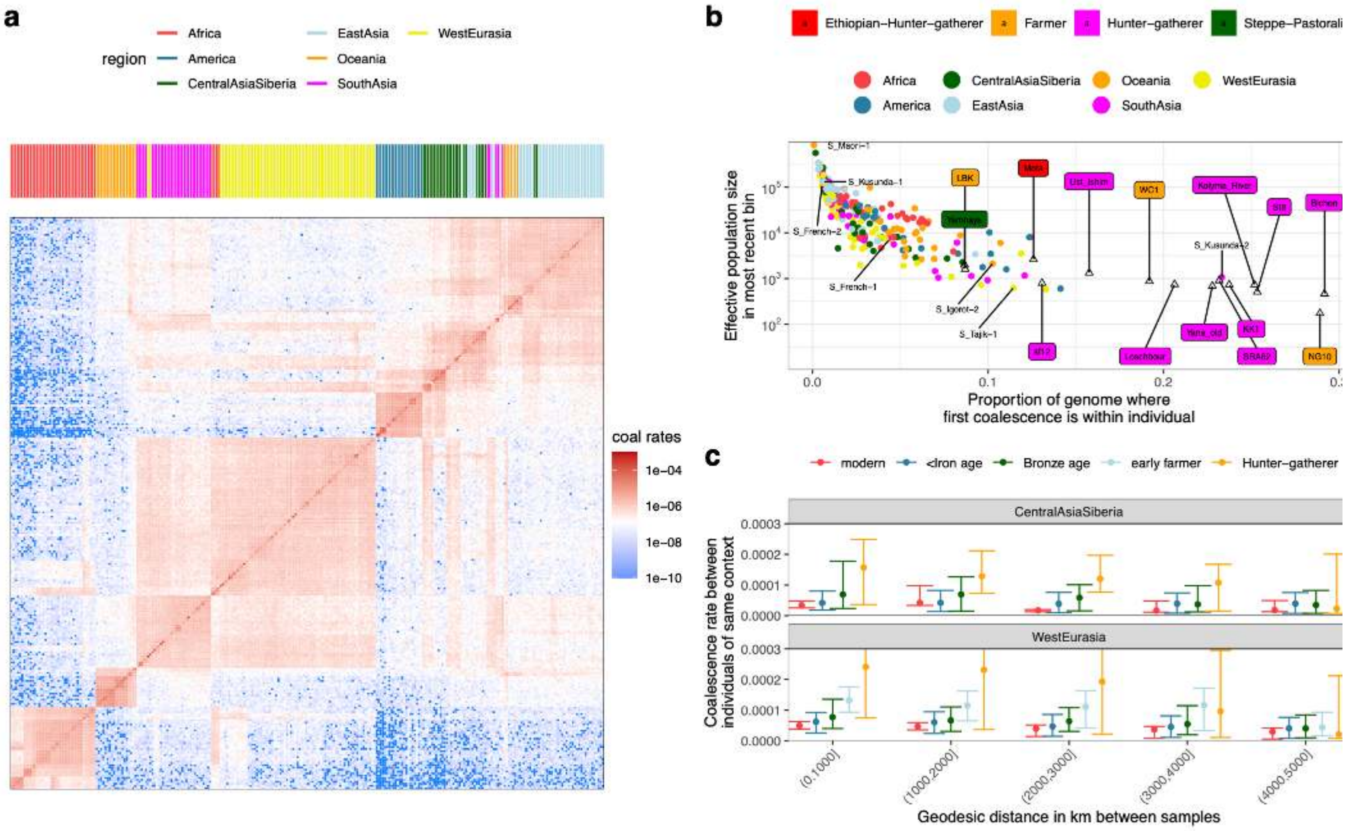

Figure 5

a, *Relate*-inferred coalescence rates between SGDP individuals in the most recent epoch (0 – 1,000 YBP). **b,** Within individual effective population sizes in the most recent epoch vs. the proportion of the genome where the first coalescence occurs within the individual. All coalescence rates are calculated using *Relate* trees. **c**, *Colate*-inferred coalescence rates in the most recent epoch (<15k YBP) for pairs of samples grouped by geographic distance and time period. Circles indicate median, error bars show the 2.5% and 97.5% percentiles, respectively.

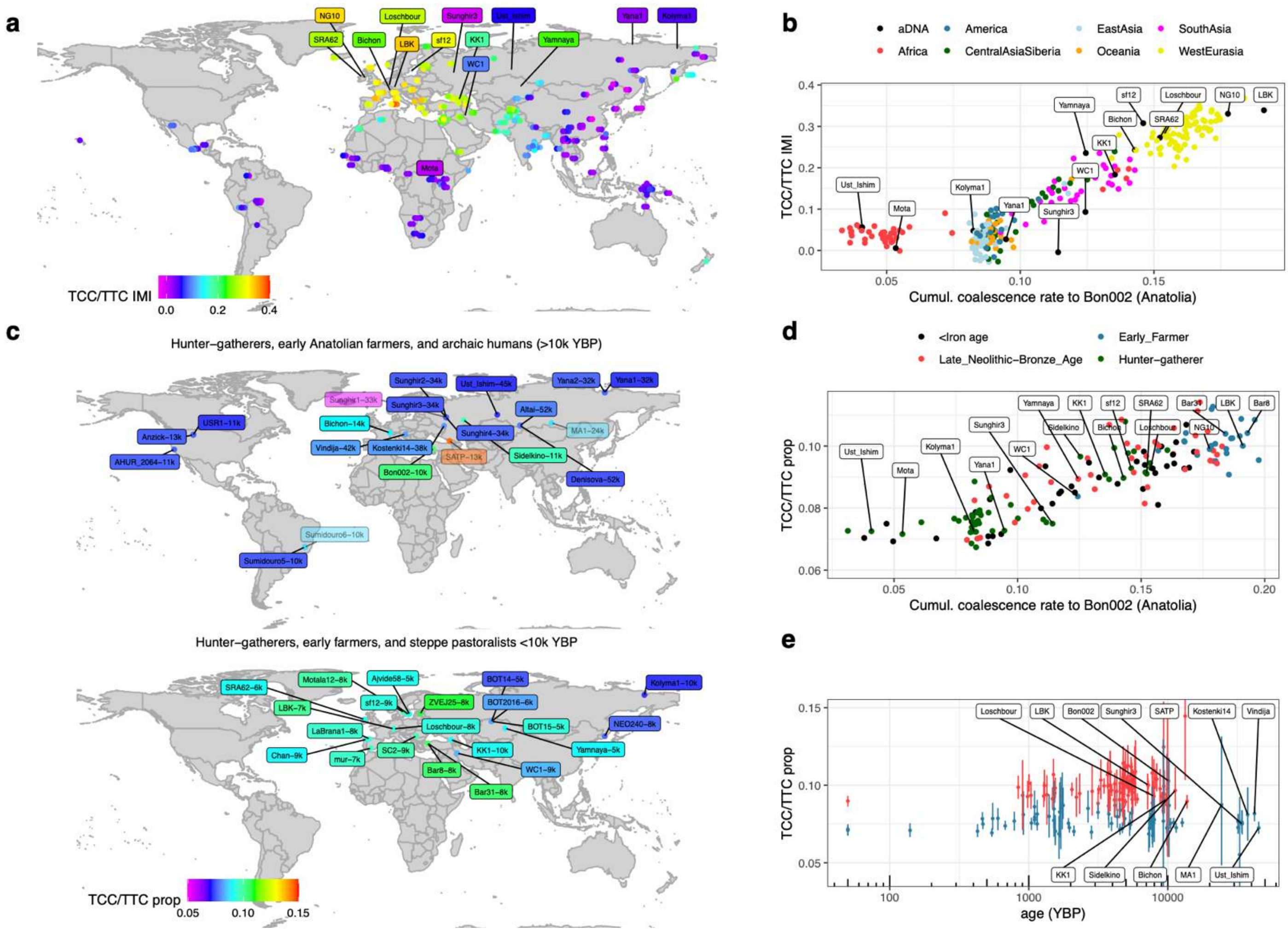

Figure 6

a, Map showing the strength of the TCC/TTC mutation rate signature, quantified by calculating the "integrated mutation intensity" (IMI) of the TCC/TTC mutation rate (**Methods**). Circles correspond to present-day individuals in the SGDP data, ancient individuals are labelled. **b)** TCC/TTC IMI plotted against the *Colate*-inferred coalescence rates to Bon002, a 10k-year-old individual from Anatolia, integrated between 14k – 50k YBP. Circles correspond to SGDP samples, labels to ancients. **c,** Map showing the TCC/TTC mutation rate signature in lower coverage ancients, quantified as the proportion of sites that are TCC/TTC relative to other C/T transitions excluding those in CpG contexts (**Methods**). Top shows a subset of samples >10k years old, bottom shows samples <10k years old (see Supplementary Figure 13 for further samples). Samples of <2x mean coverage are shown with increased transparency and number following sample ID shows sample age. **d,** Proportion of TCC/TTC sites plotted against coalescence rates to Bon002, integrated between 14k – 50k YBP. All points correspond to ancients, colour indicates their age. **e**, Proportion of TCC/TTC sites plotted against sample age. Confidence intervals are obtained using a block bootstrap. Samples are coloured using a k-means clustering (k = 2). In **c,d,e,** samples are >2x mean coverage, except for those >10k years old where we included samples >1x mean coverage.

3.4 Effective population sizes increased from Mesolithic Europe to the present

Effective population sizes calculated within an individual quantify diversity and relatedness of parental genomes. By focussing on the very recent past (<1000 years), we observe a broad spectrum of recent within-individual effective population sizes in SGDP individuals ranging from a few thousand to hundreds of thousands not limited to particular geographical groups (Figure 5**a,** Supplementary Figures 8) and correlating well between *Relate* and *Colate* (Supplementary Figures 9). Haplotypes of individuals with small recent effective population sizes coalesce with each other before coalescing with any other sample for larger proportions of the genome (Figure 5**b**), indicative of longer runs of homozygosity (ROH) in these individuals (Supplementary Figure 10). While global patterns are comparable to previously reported heterozygosity estimates (Mallick et al. 2016), the differences among particular individuals are more pronounced in our analysis, which focuses on very recent time.

Small recent effective population sizes are also observed in the high coverage ancient genomes and are most pronounced in European Mesolithic HGs, who also tend to coalesce with themselves for larger proportions of the genome. However this may at least in part be driven by increased divergence from other samples, in addition to ROH (Figure 5**b**). The smallest recent effective population size is observed for the NG10 individual, a 5,200-year-old Neolithic individual buried in a Megalithic tomb in Ireland, who was previously identified to be the son of first-degree relatives (Cassidy et al. 2020). We next compared coalescence rates across individuals at increasing geographic distances within Europe, and within Central Asia, in each time period, including only modern individuals within 500km of an ancient sample (Figure 5**c**). At shorter distances we observe a clear trend for smaller coalescence rates (larger effective population sizes) towards the present, suggesting strongly increasing local population sizes. At larger distances the relationship is non-monotonic, with coalescence rates not decreasing consistently, implying a trend of increasing migration, countering the larger population sizes. Finally, we see a trend of decreasing similarity with distance, implying local population structure at all times, with the interesting exception of samples more recent than the beginning of the Iron age (yet not modern) in Europe. More widespread sampling is needed to understand this pattern, although this period does overlap e.g., increased mobility during the Roman Empire and the following “migration age” in Europe characterized by widespread movements of peoples (Martiniano et al. 2016).

3.5 Elevation in TCC to TTC mutation rate is present in Mesolithic HGs and Neolithic farmers

The triplet TCC has seen a remarkable increase in mutation rates towards TTC in humans, first identified by (Harris 2015). This signature has no known cause to date and appears strongest in Europeans and weaker in South Asians. It was previously estimated to have started around 15k – 20k YBP, and its driver is most likely absent in present-day individuals (Harris and Pritchard 2017; Speidel et al. 2019), although there is considerable uncertainty about this estimate – for example, a recent study dates the onset to up to ~80k YBP depending on the demographic history used (DeWitt et al. 2020). One study previously quantified the signal in an early farmer (LBK) and Western HG (Loschbour), suggesting that both carried the signal, while the signal was missing in Ust'-Ishim, Neanderthals, and Denisovans (Mathieson and Reich 2017).

We first inferred the rate through time at which TCC mutates towards TTC in every individual built into our genealogy of moderns and ancients, after excluding singletons, and then quantified signal strength by calculating the "integrated mutation intensity" (IMI) which quantifies the area under the mutation rate curve (**Methods**). Among SGDP individuals, the quantified signal varies and is strongest in Southern Europeans such as Sardinians, who are known to have an increased affinity to early Neolithic farmers (Figure 6**a,** Supplementary Figure 11). Among the high-coverage ancients built into our *Relate* genealogies, we observe the signature in Mesolithic HGs, as well as in Neolithic and Bronze age samples, including the Yamnaya (Figure 6**a**), but infer it to be weaker in HGs and strongest in Neolithic farmers. The signal is absent in an Ethiopian HG, as expected, as well as in both the 45,000 year old Ust'-Ishim sample and the 34,000 year-old Sunghir3 sample (Figure 6**a**).

To quantify the signal in individuals of lower coverage, we calculate the proportion of TCC/TTC mutations relative to C/T transitions in each individual, restricting to mutations ascertained in SGDP samples, of at least 4x coverage in the ancient, and dated by *Relate* to be <100k YBP (**Methods**). We confirm that signal strength is highly correlated (96%) to our IMI estimate for the high-coverage samples built into our *Relate* genealogy, where both estimates are available (Supplementary Figure 12). We do not observe the signal in Neanderthals (Prüfer et al. 2014; Prüfer et al. 2017) or Denisovans (Meyer et al. 2012), consistent with (Mathieson and Reich 2017). The signal appears already widespread in the Late Upper Paleolithic, as it is carried by Bichon (Western HG; 13.7k YBP), by Sidelkino (Eastern HG; 11k YBP), by SATP (Caucasus HG; 13k YBP), and Bon002 (Early Neolithic Anatolian; 10k YBP) (Figure 6**c,** Supplementary Figure 13).

We note that the Caucasus HG SATP has a strong signal, however confidence intervals are large due to its lower coverage and this estimate may therefore be somewhat unreliable, although it seems clear that this individual carried the signal, which is also present in a later higher coverage Caucasus HG (KK1; 8k YBP). The Mal'ta individual (MA1) (Raghavan et al. 2014) has a similarly large confidence interval but may not have been a carrier of this signal. A 9,000 year-old Iranian farmer, WC1 (Broushaki et al. 2016), who can be modelled as a mixture of a "basal Eurasian" and Mal'ta-like ancestry, and who is not closely related to Anatolian farmers, likely only carried the signal weakly, if at all. Interestingly, Chan, a 9000-year-old Iberian HG (Olalde et al. 2019) who has little ancestry related to Western HGs such as Bichon, and instead increased affinity to HGs predating these in Europe, has the weakest signal among all Mesolithic Europeans.

Already 10,000 years ago, the signal appears weaker in Western HGs compared to the Anatolian genome, who is among the strongest carriers of this signal (similar strength to later Neolithic individuals and present-day Sardinians) (Figure 6**e**), suggesting that the driver of this mutation rate change, which may have been of genetic or environmental nature, was already extinct by the Holocene. Eastern HGs have a slightly elevated signal compared to Western HGs. Overall, this provides direct support for previous analyses based on modern-day genomes that found a reduction of the TCC/TTC pulse to normal levels in the last 10-15k years (Speidel et al. 2019; DeWitt et al. 2020) and would imply that excess TCC/TTC mutations were subsequently passed on only through shared ancestry. Strikingly, the strength of the TCC/TTC signal shows a remarkable correlation with recent coalescence rates to the 10k-year-old Anatolian individual (96% using IMI for SGDP non-Africans and 13 high-coverage ancients, 71% using TCC/TTC proportion for ancients) (Figure 6**b, d**), and does not correlate as well with coalescence rates to any other HG group for whom we have data (88% or 58% with Caucasus HGs (SATP), 83% or 53% with Scandinavian HGs (sf12), 76% or 37% with Eastern HGs (Sidelkino), 73% or 53% with Western HGs (Bichon), where first number uses IMI, second number uses TCC/TTC proportion) (Supplementary Figures 14). We therefore hypothesise that the signal spread through ancestors of this Anatolian individual across Europe before the arrival of farming, and subsequently arrived in Europe for a second time with Neolithic farmers.

The genetic relationship among West Eurasian HG groups in the Late Paleolithic is not fully understood and, to the best of our knowledge, current models do not include a clear source group contributing widely across these HG groups, while able to explain the strong correlation to ancestry from Anatolia. One potential source are ancestors of the Dzudzuana Cave individuals, a group inhabiting the Caucasus ~26k years ago (Lazaridis et al. 2018), from whom Anatolians are thought to derive the majority of their ancestry. This ancestry is present to a lesser extent in Caucasus HG and is even further diluted in Iranian Early Farmers. Dzudzuana-related populations may also have contributed ancestry to Eastern and Scandinavian HGs before the spread of farming. The Dzudzuana individuals have a pre-LGM common ancestor with Western HGs, including Bichon, however, placing the signal on this common ancestor lineage does not immediately explain the signal strength difference and correlation to shared ancestry with Anatolia. Two potential explanations include: the mutation rate elevation occurred in a Dzudzuana-like joint ancestor of Anatolian farmers and Western HGs, with subsequent dilution of the signal in Western HGs from an ancestry not closely related to Anatolia. Another possibility is that the mutation rate elevation occurred in a group more specific to Anatolia and that the signal spread during the Bølling-Allerød interstadial, a brief warming following the last glacial maximum, during which Western HGs spread across Europe replacing earlier HG groups and which may have introduced gene-flow from the Near East into Europe (Fu et al. 2016).

We note that while the cause of this mutation rate elevation remains uncertain, our results would fit well with a genetic cause within a specific ancient population (for example a mutation in some repair protein, transiently present). If, alternatively, the cause is environmental, it appears highly localised in both time and place, and this seems potentially harder to explain.

4 Discussion

The last decade has seen an explosion in the number of sequenced ancient genomes, uncovering remarkable stories of population replacements and admixture that are associated with dramatic shifts in lifestyle arounds the world (Skoglund and Mathieson 2018). While ancient genomes are still typically available in smaller numbers and lower quality compared to genomes of present-day people, they are uniquely valuable in providing direct insight into the genetic makeup of our ancestors. We have extended the *Relate* method for inference of genome-wide genealogies to work with ancient genomes and introduced a new method, *Colate*, for inference of coalescence rates for low-coverage unphased genomes. Together, these tools enable us to harness the power of genealogy-based analyses on a wider range of samples, including those of lower quality, which were previously inaccessible.

We demonstrated, using 278 moderns of the SGDP data set, 14 high-coverage, and 430 lower-coverage ancients, that *Relate* and *Colate* can uncover dynamic population histories and evolution in the processes that drive genetic variation. The extent to which directional gene-flow occurred from groups related to ancient Anatolia into European HGs predating the spread of farming in Europe has remained controversial. We have provided two further lines of evidence that such gene-flow existed, first using coalescence rates of lineages recently coalesced between Anatolia and HGs. The TCC/TTC mutation rate elevation in all these ancient groups, and its strong correlation to inferred recent shared ancestry with Anatolia, offers complementary support that the shared ancestry detected by Colate indeed reflects recent gene exchange, given the age distribution of samples showing this mutational phenomenon.

Future avenues of research may include using genealogies for parametric inference of population histories and admixture, inspired by approaches based on site-frequency spectra (Excoffier et al. 2013; Terhorst et al. 2017) and F-statistics (Patterson et al. 2012; Peter 2016; Ralph et al. 2020). Coalescence rates can be interpreted as a function of gene flow (or the lack thereof); for instance, (Wang et al. 2020) have recently developed a method that infers migration rates through time given pairwise coalescence rate estimates. Genealogies of modern individuals have proven to be powerful in quantifying positive selection (Speidel et al. 2019; Stern et al. 2019; Stern et al. 2021) and genealogies including ancient genomes should further boost power.

While *Colate* has made it possible to leverage genealogies for the study of low-coverage genomes possible, we ideally would like to incorporate such genomes directly into genealogical trees. This is currently not possible, however recent work building on the tsinfer methodology (Kelleher et al. 2019) provides an alternative approach that constrains the age of ancestral haplotypes using low-coverage ancient genomes to infer genome-wide genealogies for higher-quality phased sequences (incl. ancients and moderns) (Wohns et al. 2021). A possibility for making lower coverage ancient genomes, or indeed hybrid capture array data, accessible to these methods is imputation (Gamba et al. 2014; Hui et al. 2020; Rubinacci et al. 2020). A potential concern is that imputation may introduce biases, particularly in ancient genomes with ancestries that are not well reflected in modern groups. These biases are often difficult to assess. Because *Colate* does not require imputation, we expect that it will be a useful tool to investigate such biases in future.

5 Methods

5.1 *Colate*

Coalescence rates are inferred by attempting to maximise the following likelihood using an expectation-maximisation (EM) algorithm. For any derived mutation carried by a reference chromosome j, we ask whether this mutation is shared by the target chromosome i, which we denote by an indicator variable $S_{\ell ij}$ (ℓ indexing SNPs). We multiply across SNPs, such that no phase information is required to compute the likelihood. To obtain coalescence rates between groups of individuals, we also multiply the likelihood across homologous chromosomes in both the target and reference groups. To calculate within-individual coalescence rates using genotypes, the method assigns one allele to each category, at random at every SNP. When input is specified in BAM format (as reference-aligned reads), we multiply across reads. The maximum likelihood estimate is then given by $\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{\ell} \prod_{i,j} P\left(S_{\ell ij} \mid a_{\ell}, \boldsymbol{\theta}\right)$, where $\boldsymbol{\theta}$ denotes piecewise-constant coalescence rates and a_{ℓ} is the age of the ℓth mutation, which we have to integrate out in practice.

To integrate out mutation age, we assume neutrality of every mutation, implying that its age is uniformly distributed on the branch onto which it maps. The EM algorithm requires us to integrate out mutation age conditional on sharing or not sharing between target and reference chromosomes. This theoretically implies a deviation from the uniform distribution. This deviation is strongest for mutations that are singletons in the genealogy used to date these mutations and are shared between sequences in the target and reference chromosome sets. In this case, knowledge of sharing implies that the mutation is older than the coalescence time of the target, reference, and corresponding individual in the genealogy, biasing mutation age upwards compared to a uniform distribution (Supplementary Figure 15). We use an empirical approach to sample mutation ages for these shared singletons and use the uniform distribution for all other mutations in practise, which we demonstrate is a reasonable approximation (SI). Moreover, we note that the *Colate* approach requires the inclusion of sites fixed and derived in all samples used for inferring the genealogy, as the additional reference and target samples can, in theory, coalesce into the root branch. To obtain an approximate upper bound on the age of such mutations, we fix the time to the most recent common ancestor (TMRCA) to an outgroup (10M YBP for human-chimpanzee in this study).

We bin mutation ages into a discrete time grid to reduce computation time of the EM algorithm. As a result, the algorithm only requires the number of shared and not-shared mutations in each time grid as input; compilation of this input data is linear in sample size and number of mutations. Once in this form, the input data to the EM algorithm, and hence the computation time of the EM algorithm, is independent of sample size or the number of mutations.

5.2 Simulations

To evaluate *Relate* and *Colate*, we used stdpopsim to simulate genomes with different demographic histories (Adrion et al. 2020) and hotspot recombination rates. For *Colate*, we additionally require an outgroup to determine mutations that are fixed in all samples. Instead of simulating an outgroup explicitly, we fixed the time to the most recent common ancestor (TMRCA) t_{out} to the outgroup ($t_{out} = 10$M years in our simulations), and sampled the number of fixed mutations in any given region as a Poisson distributed random variable with mean $\mu l(t_{out} - t_{sample})$, where μ is the per base per generation mutation rate, t_{sample} is the TMRCA of the sample in this region and l is the number of base-pairs in this region. If t_{sample} was greater than t_{out}, we sampled no fixed mutations. We then chose the base-pair positions of these fixed mutations uniformly at random with replacement within the corresponding region. For simplicity, we assumed a two-state mutation model, such that any occasional repeat mutation at one genomic site return to the original state.

Supplementary Figure 2 shows the performance on a zigzag history (Schiffels and Durbin 2014), demonstrating near perfect recovery of coalescence rates when using true mutation ages in *Colate*, and high accuracy when mutation ages are sampled given a genealogy; the discrepancy highlights that our sampling distribution of mutation age given a genealogy (**Methods**, Supplementary Information) is reasonable but not exact.

We also simulated data under a multi-population model of ancient Eurasia, previously fitted using real human genomes (Kamm et al. 2020), using the stdpopsim package. We simulated 200 haploid sequences in each of three modern human groups (Mbuti, Sardinian, Han), as well as four ancient Eurasians (LBK, Loschbour, Ust'-Ishim, MA1) and a Neanderthal (two haploid sequences in each group) (Figure 2**a**, Supplementary Figure 3). From this simulation, we obtained true genealogical trees and inferred *Relate* trees for all samples. In addition, we inferred a separate set of *Relate* trees using only the three modern human groups (Mbuti, Sardinian, Han), and used these to date mutations for *Colate*.

Colate recovered within and across group coalescence rates accurately compared to the corresponding direct MLEs calculated on true or Relate-inferred trees (Figure 2**a**).

Relate-inferred coalescence rates show that ancients coalesce with other individuals at the expected rates and in the correct order on average, as can be seen with MA1, for instance, who is inferred to have a shared history with LBK, Loschbour, and Sardinians, but from a population that splits off from Han around 50k years ago. In particular, these coalescence rates clearly captured the admixture from Neanderthals into an ancestral Eurasian lineage, as well as more recent genetic structure, such as separation of the Loschbour HG and early farmer lineages, represented by LBK. We observed a closer affinity of the Loschbour HG to modern-day Sardinians, compared to LBK, consistent with modern Sardinians being an admixture of HG and farmer ancestry in this simulation.

One case for which *Colate* performed less well compared to direct MLEs obtained from *Relate* trees is in inferring the cross-coalescence rates between Neanderthals and Mbuti, calculated by assigning the Neanderthal as reference and Mbuti as target. This is because the genealogy used to date mutations can provide dates only for variants segregating in the three modern groups. Therefore, the large majority of those Neanderthal sites that mutated more recently than the Neanderthal-Mbuti split cannot be used for inference. In this case, it would instead be preferable to assign Mbuti as reference.

5.3 Evaluating *Colate* on downsampled high-coverage genomes

We evaluated the performance of *Colate* on low-coverage sequencing data, by comparing estimates obtained from downsampled BAM files (Figure 2**b**, Supplementary Figure 3). To date mutations, we constructed a genealogy containing 25 diploid samples from each of the three 1000 Genomes populations - YRI (Yobura in Ibadan, Nigeria), CEU (Northern and Central European ancestry individuals from Utah, USA), and CHB (Han Chinese from Beijing, China) (The 1000 Genomes Project Consortium 2015), downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. We then chose four 1000 Genomes samples that were not incorporated into this genealogy as target individuals (HG00096, HG00268, NA18525, NA19017) and included the remaining samples in groups YRI, CEU and CHB in the reference chromosomes set. The BAM files of these four genomes were obtained from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140203_broad_high_cov_pcr_free_validation/matching_LC_samples_bwamem/

and subsequently downsampled to a variety of reduced sequencing coverages using SAMtools v1.9 (Li et al. 2009).

Across a wide range of mean coverages, *Colate*-inferred coalescence rates remained unchanged. To obtain 95% confidence intervals, we used a block bootstrap, dividing the genome into 20Mb blocks, and resampling 100 times. Confidence intervals become wider for lower coverage sequencing data; encouragingly, we could infer meaningful coalescence rates between a target sequence of 0.01x mean coverage and the reference VCFs.

We additionally evaluated *Colate* when both target and reference samples are of low coverage by calculating the coalescence rates between LBK, a 7200 year old early European farmer, and Loschbour, a nearly 8000 year old Mesolithic Western HG (both >14x coverage) (Lazaridis et al. 2014) using a genealogy for SGDP to date mutations. We downsampled both individuals to a minimum of 0.1x mean coverage (Supplementary Figure 4). While inference of coalescence rates became challenging when both genomes are at 0.1x, estimates still appeared reasonably accurate and unbiased.

5.4 Data

5.4.1 Simons Genome Diversity Project Data

We downloaded phased haplotypes for 278 individuals from https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/PS2_multisample_public/, and rephased these jointly with high coverage ancients (Section 4.4.2) using SHAPEIT4 (Delaneau et al. 2019). We first used the 1000 Genomes Project (1000GP) reference panel (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) to phase all sites overlapping with 1000GP and then internally phased all remaining sites, while keeping the already phased sites fixed.

5.4.2 Ancient genomes data

We downloaded 430 ancient genomes for use in this study (Supplementary Table 1). All samples had a genome-wide mean coverage of 0.5x or more. We selected 14 high coverage ancient genomes (mean genomic coverage > 7.8X) for the *Relate* analysis.

For these 14 high coverage genomes (Supplementary Table 1) genotypes were called using samtools mpileup (input options: -C 50, -Q 20 and -q 20) and bcftools call --consensus-caller with indels ignored (Li 2011). A modified version of the bamCaller.py script from https://github.com/stschiff/msmc-tools was used to output variant sites. We generated a quality mask for each ancient genome, declaring only sites with at least 5X coverage and below twice the mean genomic coverage as passing.

We merged these 14 ancient genomes with the 278 Simon Genome Diversity Project samples to infer joint genealogies using *Relate*. We constructed a conservative joint mask, declaring only sites passing in all of the 14 ancients, as well as a universal mask file provided with the SGDP data set, as passing. The SGDP universal mask was obtained from https://reichdata.hms.harvard.edu/pub/datasets/sgdp/filters/all_samples/.

5.5 Joint genealogies of ancients and moderns

We inferred joint genealogies of ancients and moderns using our updated *Relate* algorithm (Supplementary Information). We used all mutations, excluding those in CpG contexts, to infer tree topologies and then restricted to transversion only for inference of branch lengths. Assuming an overall average mutation rate of 1.25×10^{-8} per base per generation and a transition to transversion ratio of 2 in humans (Ségurel et al. 2014), we therefore reduced the mutation rate for branch length inference to 4×10^{-9} per base per generation. We used a recombination map obtained from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html and realigned alleles relative to an ancestral genome obtained from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/. We otherwise used default parameters in *Relate*.

To infer branch lengths, we used a precomputed average coalescence rate estimate obtained by applying *Relate* to the 278 SGDP modern samples. To compute these coalescence rates, we jointly sampled branch lengths and effective population sizes using our updated iterative algorithm, which we show can be interpreted as an approximate EM algorithm for finding maximum likelihood coalescence rates. This approximate EM algorithm samples genealogies using *Relate* instead of integrating over all possible genealogies (see Supplementary Information Section B). To obtain a coalescence rate estimate that matches the mutation rate used for inferring the genealogy of ancients and moderns, we inferred branch lengths using transversions only and set the mutation rate to 4x10^{-9} per base per generation.

5.6 *Colate*-inferred coalescence rates for SGDP and 430 ancient samples

We inferred coalescence rates for pairs of ancient individuals using *Colate*, restricting to transversions only. For each pair of samples, when given as a VCF file, we applied the respective mask files. When a sample was given in BAM file format, we accepted a read whenever its mapping quality exceeded 30, read length exceeded 34 bps, and there were fewer than three mismatching sites compared to the reference genome. We further excluded 2 base-pairs at each end of a read and restricted our analysis to sites where at most two different alleles were observed.

To date mutations, we used a *Relate*-inferred genealogy of the SGDP dataset. As the degree of inbreeding varied across SGDP individuals (main text) and to avoid biases in mutation ages resulting from extensive inbreeding in some individuals, we selected one haploid sequence from each individual in constructing this genealogy. We jointly fitted branch lengths and coalescence rates using a mutation rate of 1.25x10^{-8} per base per generation.

5.7 Inference of directional migration

To investigate evidence for directional migration, we focus on lineages that are recently coalesced (<50k YBP) between an individual in groups A and an individual in group B. If these groups split >50k YBP, then any such lineage should exclusively come from migrants of one group to the other, or at least should in practice be highly enriched for such migrant lineages. Therefore, if migration occurred purely from group A into group B, these recently coalesced lineages can be classified as belonging to group A back in time and should behave like any other lineage in group A. The approach is expected to also be robust to earlier split times (<50k YBP), because lineages behave identically once groups have coalesced, reflected in identical coalescence rates for epochs predating the split.

We test this by calculating coalescence rates of such recently coalesced lineages to each non-African SGDP individual, integrated between 0 – 50k YBP and stored in variable y. We expect that these coalescence rate profiles differ between lineages assigned to groups A and B and we can distinguish whether a lineage belongs to either group. We fit this ($N \times 1$ vector, N being the number of non-African SGDP individuals) vector y as a mixture of coalescence rates (also integrated between 0-50k YBP) of k surrogate source individuals to non-African SGDP individuals, denoted by x_k ($N \times 1$ vectors). We use non-negative least squares, such that the coefficients β ($k \times 1$ vector) are given by

$$\hat{\beta} = \operatorname{argmin}_{\beta \geq 0} \|y - X\beta\|_2,$$

where $\| \quad \|_2$ denotes the Euclidean norm and X is a $N \times k$ matrix with columns given by x_k. We use the R function nnls to find non-linear least squares estimates and bootstrap entries of our vectors y and x_k to obtain confidence intervals.

5.8 Calculation of mutation rate

We calculated mutation rates for 76 mutation triplets (of 4x4x3/2=96 possible) in each individual, after excluding any singletons and terminal branches in our genealogy. We only considered mutation triplets that are not in a CpG context, which excludes 20 possible triplets. To remove trends shared across mutation triplets, we divided the TCC/TTC mutation rate by the average over all triplets (excl. CpG contexts) in each epoch, to obtain the mutation rate relative to the average mutation rate.

To calculate the area under the curve for the TCC/TTC mutation rate signature, we first scaled the mutation rate for this triplet in each individual by the average across triplets over the time interval [1e5,1e6] YBP (predating the emergence of this signature). We then calculated the integrated mutation intensity (IMI), which is the area under the curve between 14k to 1M years BP, where time is measured in log10 units to upweight the recent past. For samples that are older than 14k years (Ust'-Ishim, Sunghir3, and Yana1), we extrapolated the earliest value to 14k YBP. We then subtracted the equivalent value of a constant mutation rate from this IMI, such that any sample without the elevation in TCC/TTC mutation rates is expected to have an IMI of 0.

5.9 Quantifying the TCC/TTC signal in lower coverage individuals

We quantified the TCC/TTC signal in lower coverage individuals (>2x mean coverage) by restricting to sites segregating in our SGDP genealogy that we also used to date mutation in *Colate*. We additionally restricted to sites where the age of the upper coalescence event of the branch onto which the mutation maps is <100k YBP. For each sample, at any such site, we then further restricted to sites where there were at least four reads mapping and added a count towards a mutation category in that individual if at least four reads supported the derived allele. In this way, we counted the number of sites with strong evidence of being in a heterozygous or homozygous state for the derived allele. We finally calculated the proportion of such sites, relative to any C/T transitions, excluding those in CpG context. We calculated confidence intervals using a block bootstrap with block size of 10Mb.

Ascertainment of mutations in moderns may potentially downwards bias signal strength in some ancients, if these possess private TCC/TTC variants less likely to be transmitted to modern individuals compared to other transitions. This could happen for instance if close ancestors of an individual carried the driver of this mutation rate pulse generating private variants. However, regardless, we still expect this approach to find the group from which the signal spread into modern-day humans. In addition, the overall good agreement with the IMI estimates obtained from Relate genealogies of high-coverage samples (Supplementary Figure 12), where no such ascertainment is done, we believe that any such biases have only a minor effect.

5.10 Calculation of pairwise F2 statistics

We calculated F2 statistics between ancients for comparisons to matrices of pairwise coalescence rates (used in Supplementary Figure 7). To calculate F2 statistics, we first made pseudohaploid calls for each individual using "pileupcaller" (https://github.com/stschiff/sequenceTools), where we restricted to 1240k ascertained genomic sites known to be varying among present-day humans (Mathieson et al. 2015). We then merged individuals using "mergeit" (https://github.com/DReichLab/EIG). To calculate F2 statistics, we used the R package admixtools2 (https://github.com/uqrmaie1/admixtools).

Acknowledgements

We thank David Reich, Richard Durbin, Iain Mathieson, Mateja Hajdinjak, Anders Bergström, Simone Rubinacci, and Clare E. West for helpful comments, discussions, and ideas. L.S. acknowledges support provided by the Sir Henry Wellcome fellowship [220457/Z/20/Z]. L.S., G.H. and S.R.M acknowledge funding from the Wellcome Trust [200186/Z/15/Z] and S.R.M. acknowledges support provided by the Wellcome Trust Investigator Award [212284/Z/18/Z]. P.S. is supported by the Francis Crick Institute core funding (FC001595) from Cancer Research UK, the UK Medical Research Council, and the Wellcome Trust. P.S. is also supported by the European Research Council (grant no. 852558), a Wellcome Trust Investigator award (217223/Z/19/Z) and the Vallee Foundation. This research was funded in whole, or in part, by the Wellcome Trust. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author contributions

L.S. and S.R.M designed the study and developed the methods. L.C. prepared the data. All authors analysed the results and wrote the manuscript.

Software availability

Relate: https://myersgroup.github.io/relate/
Colate: https://github.com/leospeidel/Colate

References

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, et al. 2020. A community-maintained standard library of population genetic models. *Elife* 9:e54967.

Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, et al. 2015. Population genomics of Bronze Age Eurasia. *Nature* 522:167–172.

de Barros Damgaard P, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, Barjamovic G, Rasmussen S, Zacho C, Baimukhanov N, et al. 2018. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360:eaar7711.

Broushaki F, Thomas MG, Link V, López S, van Dorp L, Kirsanow K, Hofmanová Z, Diekmann Y, Cassidy LM, Díez-del-Molino D, et al. 2016. Early Neolithic genomes from the eastern Fertile Crescent. *Science* 353:499–503.

Cassidy LM, Maoldúin R, Kador T, Lynch A, Jones C, Woodman PC, Murphy E, Ramsey G, Dowd M, Noonan A, et al. 2020. A dynastic elite in monumental Neolithic society. *Nature* 582:384–388.

Cassidy LM, Martiniano R, Murphy EM, Teasdale MD, Mallory J, Hartwell B, Bradley DG. 2016. Neolithic and Bronze Age migration to Ireland and establishment of the insular atlantic genome. *Proc. Natl. Acad. Sci. U. S. A.* 113:368–373.

Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* 10:24–29.

DeWitt WS, Harris KD, Harris K. 2020. Joint nonparametric coalescent inference of mutation spectrum history and demography. *bioRxiv*:2020.06.16.153452.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet.* 9:e1003905.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, De Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514:445–449.

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A, et al. 2016. The genetic history of Ice Age Europe. *Nature* 534:200–205.

Gallego-Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, et al. 2015. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science* 350:820–822.

Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, Domboróczki L, Kovári I, Pap I, Anders

A, et al. 2014. Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* 5:1–9.

Günther T, Malmstro H, Sa F, Krzewi M, Eriksson G, Fraser M, Edlund H, Munters AR, Coutinho A, Sjo A, et al. 2018. Population genomics of Mesolithic Scandinavia : Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol.* 16:e2003703.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:1000695.

Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Der Sarkissian CSI, Brandt G, Schwarz C, Nicklisch N, et al. 2010. Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. *PLoS Biol.* 8:e1000536.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207–211.

Harris K. 2015. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U. S. A.* 112:3439–3444.

Harris K, Pritchard J. 2017. Rapid evolution of the human mutation spectrum. *Elife* 6:e24284.

Hui R, D'Atanasio E, Cassidy LM, Scheib CL, Kivisild T. 2020. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci. Rep.* 10:18542.

Jones ER, Gonzalez-Fortes G, Connell S, Siska V, Eriksson A, Martiniano R, McLaughlin RL, Gallego Llorente M, Cassidy LM, Gamba C, et al. 2015. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* 6:8912.

Kamm J, Terhorst J, Durbin R, Song YS. 2020. Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *J. Am. Stat. Assoc.* 115:1472–1487.

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. 2019. Inferring whole-genome histories in large population datasets. *Nat. Genet.* 51:1330–1338.

Kılınç GM, Omrak A, Özer F, Günther T, Büyükkarakaya AM, Bıçakçı E, Baird D, Dönertaş HM, Ghalichi A, Yaka R, et al. 2016. The Demographic Development of the First Farmers in Anatolia. *Curr. Biol.* 26:2659–2666.

Lazaridis I, Belfer-Cohen A, Mallick S, Patterson N, Cheronet O, Rohland N, Bar-Oz G, Bar-Yosef O, Jakeli N, Kvavadze E, et al. 2018. Paleolithic DNA from the Caucasus reveals core of West Eurasian ancestry. *bioRxiv*:10.1101/423079.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513:409–413.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493–496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538:201–206.

Martiniano R, Caffell A, Holst M, Hunter-Mann K, Montgomery J, Müldner G, McLaughlin RL, Teasdale MD, van Rheenen W, Veldink JH, et al. 2016. Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat. Commun.* 7:10326.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528:499–503.

Mathieson I, Reich D. 2017. Differences in the rare variant spectrum among human populations. *PLOS Genet.* 13:e1006581.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, De Filippo C, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338:222–226.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. 2008. Genes mirror geography within Europe. *Nature* 456:98–101.

Olalde I, Mallick S, Patterson N, Rohland N, Villalba-mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M, et al. 2019. The genomic history of the Iberian Peninsula over the past 8000 years. 1234:1230–1234.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192:1065–1093.

Peter BM. 2016. Admixture, population structure, and f-statistics. *Genetics* 202:1485–1501.

Prüfer K, De Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. 2017. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* 358:655–658.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43–49.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, et al. 2014. Upper palaeolithic Siberian genome reveals dual ancestry of native Americans. *Nature* 505:87–91.

Ralph P, Thornton K, Kelleher J. 2020. Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. *Genetics* 215:779–797.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* 10:e1004342.

Rubinacci S, Ribeiro DM, Hofmeister R, Delaneau O. 2020. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *bioRxiv*: 2020.04.14.040329.

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* 46:919–925.

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V, et al. 2014. Genomic structure in Europeans dating back at least 36,200 years. *Science* 346:1113–1118.

Ségurel L, Wyman MJ, Przeworski M. 2014. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* 15:47–70.

Sikora M, Pitulko V V., Sousa VC, Allentoft ME, Vinner L, Rasmussen S, Margaryan A, de Barros Damgaard P, de la Fuente C, Renaud G, et al. 2019. The population history of northeastern Siberia since the Pleistocene. *Nature* 570:182–188.

Sikora M, Seguin-Orlando A, Sousa VC, Albrechtsen A, Korneliussen T, Ko A, Rasmussen S, Dupanloup I, Nigst PR, Bosch MD, et al. 2017. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* 358:659–662.

Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, Hall P, Tambets K, Parik J, Sjögren KG, et al. 2014. Genomic diversity and admixture differs for stone-age Scandinavian foragers and farmers. *Science*344:747–750.

Skoglund P, Mathieson I. 2018. Ancient Genomics of Modern Humans: The First Decade. *Annu. Rev. Genomics Hum. Genet.* 19:381–404.

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* 51:1321–1329.

Stern AJ, Speidel L, Zaitlen NA, Nielsen R. 2021. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am. J. Hum. Genet.* 108:219–239.

Stern AJ, Wilton PR, Nielsen R. 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLOS Genet.* 15:e1008384.

Terhorst J, Kamm JA, Song YS. 2017. Robust and scalable inference of population history froth hundreds of unphased whole genomes. *Nat. Genet.* 49:303–309.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68–74.

Wang K, Mathieson I, O'Connell J, Schiffels S. 2020. Tracking human population structure through time from whole genome sequences. *PLOS Genet.* 16:e1008552.

Wohns AW, Wong Y, Jeffery B, Mallick S, Pinhasi R, Patterson N, Reich D, Kelleher J, Mcvean G. 2021. A unified genealogy of modern and ancient genomes. *bioRxiv*.