ARTICLE

Patterns of genetic connectedness between modern and medieval Estonian genomes reveal the origins of a major ancestry component of the Finnish population

Toomas Kivisild,[1,2,4,*] Lehti Saag,[2,3] Ruoyun Hui,[4,5] Simone Andrea Biagini,[1] Vasili Pankratov,[2] Eugenia D'Atanasio,[6] Luca Pagani,[2,7] Lauri Saag,[2] Siiri Rootsi,[2] Reedik Mägi,[2] Ene Metspalu,[2] Heiki Valk,[8] Martin Malve,[8] Kadri Irdt,[2] Tuuli Reisberg,[11] Anu Solnik,[11] Christiana L. Scheib,[2,4,9] Daniel N. Seidman,[10] Amy L. Williams,[10] Estonian Biobank Research Team,[2] Kristiina Tambets,[2,12] and Mait Metspalu[2,12]

Summary

The Finnish population is a unique example of a genetic isolate affected by a recent founder event. Previous studies have suggested that the ancestors of Finnic-speaking Finns and Estonians reached the circum-Baltic region by the 1st millennium BC. However, high linguistic similarity points to a more recent split of their languages. To study genetic connectedness between Finns and Estonians directly, we first assessed the efficacy of imputation of low-coverage ancient genomes by sequencing a medieval Estonian genome to high depth (23×) and evaluated the performance of its down-sampled replicas. We find that ancient genomes imputed from >0.1× coverage can be reliably used in principal-component analyses without projection. By searching for long shared allele intervals (LSAIs; similar to identity-by-descent segments) in unphased data for >143,000 present-day Estonians, 99 Finns, and 14 imputed ancient genomes from Estonia, we find unexpectedly high levels of individual connectedness between Estonians and Finns for the last eight centuries in contrast to their clear differentiation by allele frequencies. High levels of sharing of these segments between Estonians and Finns predate the demographic expansion and late settlement process of Finland. One plausible source of this extensive sharing is the 8th–10th centuries AD migration event from North Estonia to Finland that has been proposed to explain uniquely shared linguistic features between the Finnish language and the northern dialect of Estonian and shared Christianity-related loanwords from Slavic. These results suggest that LSAI detection provides a computationally tractable way to detect fine-scale structure in large cohorts.

Introduction

Evidence derived from archaeology and genome-scale studies of ancient human remains explain high genetic homogeneity across present-day Europe in a world context by massive population movements associated with Steppe ancestry in the Late Neolithic and Early Bronze Age.[1] Underneath this overarching homogeneity of allele frequencies, substantial regional differences can be revealed through the study of long identical-by-descent (IBD) segments that are sensitive to signals of regional mating patterns within the last millennia.[2] While ancient DNA work has become pivotal for addressing questions about the genetic ancestry in European prehistory, the use of IBD-based methods has been limited so far because of the fact that these require good-quality genotype calls, which can be made directly only from high-quality data. A study of a late-medieval 11.3× genome from Barcelona[3] showed, intriguingly, an excess of IBD sharing locally with the present-day Spanish population, highlighting the potential of IBD sharing measures to be informative in ancient DNA analyses in historical time depths. However, most ancient genomes that are currently available have low coverage and are routinely assessed via haploid genotype calls. Yet, accurate imputation methods[4–6] have been shown to enable the recovery of usable diploid genotype calls from ancient DNA,[7–9] including from samples with as low as 0.1× coverage data with accuracy of common variants > 0.95.[10] In parallel, fast methods for IBD estimation from tens to hundreds of thousands of individuals have been recently developed for phased[11–14] and unphased[15,16] genomic data along with scalable clustering methods for the detection of fine-scale community structure.[17,18]

Late Bronze and Early Iron Age migrations have been argued to be responsible for the spread of Finnic languages together with a minor Siberian genetic component (Figure 1D) in the circum-Baltic region;[19,20,21] however, it has been less clear how much gene flow and contact over the Gulf of Finland has occurred in the last 2,000 years (Figures 1E and 1F). Linguistic studies have suggested

[1]Department of Human Genetics, KU Leuven, Leuven 3000, Belgium; [2]Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu 51010, Estonia; [3]Research Department of Genetics, Evolution, and Environment, University College London, London WC1E 6BT, UK; [4]McDonald Institute for Archaeological Research, University of Cambridge, Cambridge CB2 3ER, UK; [5]The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK; [6]Instituto di Biologia e Patologia Molecolari, Consiglio Nazionale delle Ricerche, Rome, Italy; [7]Department of Biology, University of Padova, 35131 Padova, Italy; [8]Department of Archaeology, Institute of History and Archaeology, University of Tartu, Tartu 51014, Estonia; [9]St John's College, University of Cambridge, Cambridge CB2 1TP, UK; [10]Department of Computational Biology, Cornell University, Ithaca, NY 14853, USA; [11]Core Facility, Institute of Genomics, University of Tartu, Tartu 51010, Estonia
[12]These authors contributed equally
*Correspondence: toomas.kivisild@kuleuven.be
https://doi.org/10.1016/j.ajhg.2021.07.012.

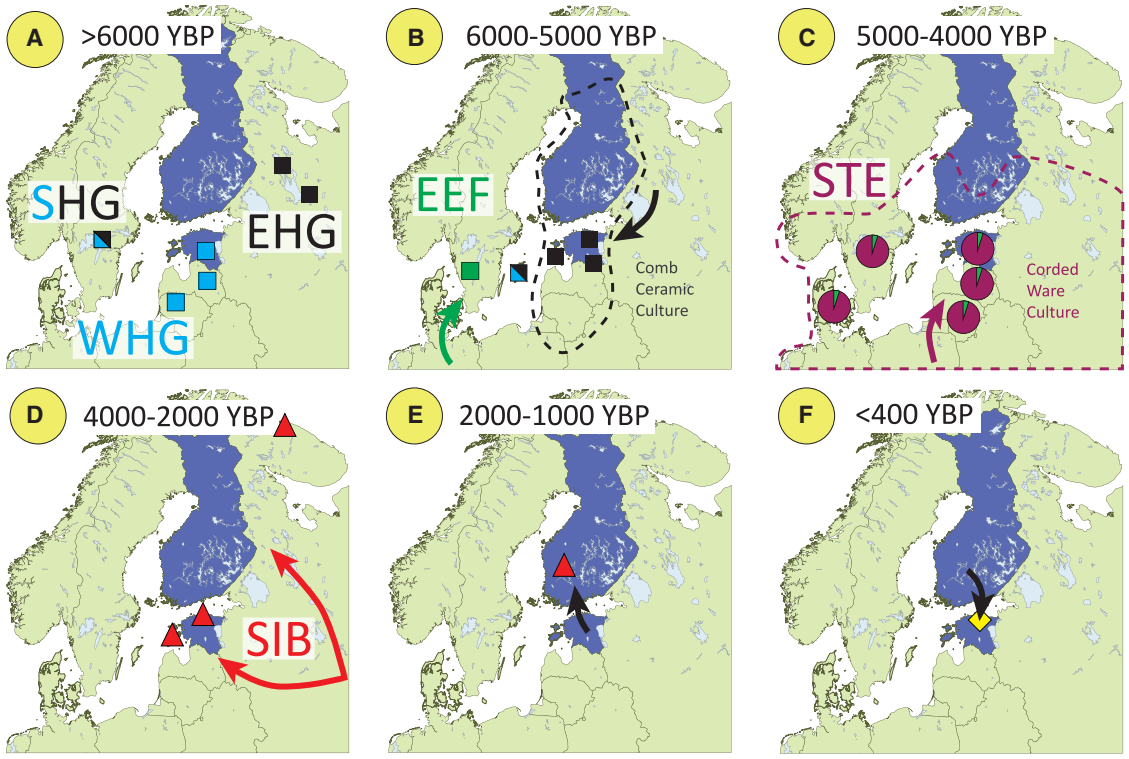


Figure 1. Simplified summary of population history of Estonia and Finland
(A–C) Initial settlement by EHG (eastern hunter-gatherers), WHG (western hunter-gatherers), and SHG (Scandinavian hunter-gatherers) (the locations of sites that have produced ancient DNA evidence are shown with squares), followed by inflow of EEF (early European farmer) and STE (Steppe) ancestry[19,31–35] (site locations and proportions of Steppe ancestry is shown with pie charts).
(D) The first appearance of SIB ("Siberian-like") autosomal and Y chromosome (haplogroup N) ancestry[19] dates back to the Late Bronze-Early Iron Age (sites locations shown with triangles) and corresponds to the likely arrival time of the Finnic languages in the region.[22]
(E) The split time of Estonian and Finnish languages has been estimated by a range of linguists at 2,000–1,000 years before present (YBP); Honkola et al.'s[22] analyses place the split time at 800 AD.
(F) Some Finnish settlements in Northeast Estonia (highlighted with a yellow diamond) date back to 17th—18th centuries.

that the differentiation of the Finnic from Finno-Volgaic languages dates back to 3,000–4,000 years ago.[22,23,24] Numerous Baltic loan words in Finnic and archaeological evidence of metal work and proximity of fortified settlements point to extensive local contacts between the Finnic and Baltic speakers in the Late Bronze and Early Iron Ages,[25] while the divergence of the Finnic and split of Estonian and Finnish may have occurred more recently between 1,000 and 2,000 years ago.[22,23] The time gap between these two split dates means that the divergence of Finnic languages likely postdates the first arrival of Finnic languages in the region. Numerous Slavic loan words in Finnish related to the spread of Christianity,[24] the similarities between North Estonian and Finnish, and the lack of record for historically attested migration events to Finland from the south since the 12th century point to a possible prehistoric migration event from North Estonia to Finland after the second wave of Slavic expansion (8th–10th centuries) potentially related to the intensification of agriculture in the region.[22,26,27] The origin of the modern Finnish population with its unique "disease heritage" has been ascribed to founder events and population range expansions, from relatively small coastal distribution of ~50,000 people to more than 5 million, within the last millennium.[28,29] Further significant founder events most likely postdate the reforms introduced by the Swedish King Gustav Vasa in the 16th century.[30]

Finns and Estonians can be clearly distinguished in genetic distance-based analyses of modern genomes.[36] Both Estonia and Finland show internally high levels of substructure,[36–38] which in the case of Finland, seems to reflect geographic divisions and founder events during the late settlement process and its long-term isolation in the last 100 generations. There is no historical record for the last eight centuries of significant migration events across the Gulf of Finland apart from the accounts of Finnish settlements in Northeast Estonia in the 17th and 18th centuries, which could account for local patterns of Finnish IBD sharing in Northeast Estonia.[38] However, the analyses of modern genomes cannot offer conclusive answers about the time depth and directionality of migration events that have caused regional and inter-regional patterns of genetic differentiation and similarity between Estonians and Finns. In this study, we focus on the potential of applying IBD-based methods on ancient genomic sequence data to address these questions. Some of our key results leverage an unphased IBD detector,[16] which is typically viewed as reliable for segments $\geq$7 cM long, yet

we find meaningful genetic signals by using shorter segments (>5 cM). Because stretches of shared alleles in an unphased context at these lengths are unlikely to always correspond to shared haplotypes,[39] we refer to them as long shared allele intervals (LSAIs).

Material and methods

Present-day populations: Estonian Biobank data merged with 1000 Genomes Project European data

Illumina Global Screening Array data for 707,385 SNVs genotyped in 150,415 individuals of the Estonian Biobank (EstBB) were merged with the 1000 Genomes Project (1000 GP) Phase 3[40] data. After applying --maf 0.05, --geno 0.005, and --mind 0.01 filters in PLINK-1.9.0[41] data for 143,774 EstBB individuals (representing >10% of the total population of Estonia), we retained 503 individuals of European ancestry (CEU, Utah residents with Northern and Western European ancestry; GBR, British in England and Scotland; FIN, Finnish in Finland; IBS, Iberian populations in Spain; and TSI, Toscani in Italy) from the 1000 GP and 254,325 overlapping SNVs with genetic map coordinates (build 37) as an input for downstream analyses.

Ancient DNA extraction and sequencing

As part of this study, DNA was extracted from medieval human remains of two individuals from Estonia: TPM003 from Tartu Püha Maarja Kirik, Tartu County and TUD001 from Tudulinna, Ida-Viru County. In addition, we analyzed a medieval tooth sample, PSN177, from the cemetery of the Hospital of St John, Cambridge, UK as a control to test the effect of local Estonian reference panel on imputation results and their effect on downstream population genetic analyses. The teeth used for DNA extraction were obtained with relevant institutional permissions from the Institute of History and Archaeology, University of Tartu and the Cambridge Archaeological Unit, Department of Archaeology, University of Cambridge, which excavated the remains from the cemetery of the Hospital of St John on behalf of St John's College . All laboratory work was performed in dedicated ancient DNA laboratories of the Institute of Genomics, University of Tartu and the Department of Archaeology, University of Cambridge. The library quantification and sequencing were performed at the Institute of Genomics Core Facility, University of Tartu.

For extraction, we broke off or cut off apical tooth roots by using a drill and used them whole to avoid heat damage during powdering with a drill and to reduce the risk of cross-contamination between samples. Contaminants were removed from the surface of tooth roots by soaking in 6% bleach, rinsing with milli-Q water (Millipore) and 70% ethanol, and drying under a UV light. Next, we added EDTA and proteinase K and left the samples to digest on a rotating mixer at 20°C for 72 h to compensate for the smaller surface area of the whole root compared to powder. The DNA solution was concentrated to 250 μL (Vivaspin Turbo 15, 30,000 molecular weight cut-off [MWCO] polyethersulfone [PES], Sartorius) and purified in large volume columns (High Pure Viral Nucleic Acid Large Volume Kit, Roche) with the MinElute PCR Purification Kit (QIAGEN).

We built sequencing libraries by using NEBNext DNA Library Prep Master Mix Set for 454 (E6070, New England Biolabs) and Illumina-specific adaptors[42] following established protocols.[19,42] The samples were purified between steps with the MinElute PCR Purification Kit (QIAGEN). The libraries were amplified and both the indexed and universal primer (NEBNext Multiplex Oligos for Illumina, New England Biolabs) were added by PCR with HGS Diamond Taq DNA polymerase (Eurogentec). We implemented three verification steps to make sure library preparation was successful and to measure the concentration of double-stranded DNA sequencing libraries—fluorometric quantitation (Qubit, Thermo Fisher Scientific), parallel capillary electrophoresis (Fragment Analyzer, Agilent Technologies), and qPCR.

We sequenced DNA by using the Illumina NextSeq 500 platform with the 75 bp single-end method. First, we multiplexed samples to gain low-coverage data. Later, we generated an additional four full runs of data for TPM003 to increase coverage.

Ancient sequence data processing and authentication

Before mapping, the adaptor sequences and poly-G tails were cut from the ends of DNA sequences via cutadapt 1.11.[43] We removed sequences shorter than 28 bp to avoid random mapping of sequences from other species. The sequences were mapped to reference sequence GRCh37 (hs37d5) with the Burrows-Wheeler Aligner (BWA 0.7.12)[44] algorithm mem with re-seeding disabled. After mapping, the sequences were converted to BAM format and only sequences that mapped to the human genome were kept with samtools 1.3.[45] Next, data from all flow cell lanes for the same sample were merged and duplicates were removed with picard 2.12. Indels were realigned with GATK 3.5,[46] and reads with mapping quality under 10 were filtered out with samtools 1.3.

Because of post-mortem degradation, ancient DNA can be distinguished from modern DNA by shorter fragments and a high frequency of cytosine deamination at the ends of sequences. We used the program mapDamage2.0[47] to estimate the frequency of deamination damage with results for the three newly reported genomes shown in Figure S1. mtDNA contamination was estimated with contammix.[48] This included calling an mtDNA consensus sequence based on reads with mapping quality of at least 30 and positions with at least 5× coverage, aligning the consensus with a panel of 311 human mtDNA sequences, mapping the mtDNA reads to the consensus sequence, and running contamMix 1.0-10 with the reads mapping to the consensus and the 312 aligned mtDNA sequences with the option trimBases to trim seven bases from the ends of reads. For male individuals, X chromosome contamination was also estimated via the two methods in the script contamination.R incorporated in ANGSD.[49]

Detailed summary of the sequence data of all the 14 ancient samples from Estonia used in this study, including 12 published genomes, and the PSN177 genome from Cambridge is provided in Table S1.

Genotype calling of the high-coverage genome

The genotype calls of the high-coverage TPM003 genome were determined with GATK 3.5 HaplotypeCaller[50] with Build37 reference and --genotyping_mode GENOTYPE_GIVEN_ALLELES, --output_mode EMIT_ALL_SITES, and --alleles variant.list options. In total, 12.6 million SNVs that had minor allele frequency (MAF) higher than 0.1% in a subset of 2,076 high-coverage whole-genome sequences[51] were used in variant.list. Called variants were filtered with genotype quality (GQ) > 30, read depth (DP) > 10, and genotype probability (GP) > 0.99. Details of the down-sampling of the high-coverage genome and the imputation

of the low-coverage replicas are given in the supplemental methods.

Principal-component analyses

We used FlashPCA2[52] to perform principal-component analysis (PCA) on high-coverage and imputed ancient genomes in the context of 69,218 EstBB samples and 503 Europeans (CEU, GBR, FIN, IBS, and TSI) from the 1000 GP data.[40] After merging genotype data of 15 ancient (including 14 Estonian samples and one British sample) and 69,713 modern individuals, we thinned the data by excluding variants in linkage disequilibrium with the PLINK[41] --indep-pairwise 1000 50 0.5 option and excluded the recommended[52] range of likely non-neutral regions with --exclude range exclusion_regions_hg19.txt. After thinning, 153,813 variants remained available for PCA, which was performed with default settings of FlashPCA2.

We assessed the performance of imputed ancient genomes by comparing the placement of TPM003 high coverage (23×) and its down-sampled (to 0.1×) and imputed replicas in PCA performed with FlashPCA2 and smartpca in analyses together with a sub-sample of 1,040 modern genomes (including 503 Europeans of the 1000 GP and 537 EstBB samples). We confirmed first that the two methods produce highly correlated PC1 ($r = 0.999997$) and PC2 ($r = 0.999979$) values and performed further analyses including the projections of the five 0.1× replicas of TPM003, which were haploid-called with smartpca that offers the option of projection with the least-squares equations. We observe minor shifts between the position of projected and unprojected copies of TPM003 relative to their most proximate neighbors in the EstBB data on the plot (Figure S2). Similar shifts were observed in additional analyses with different sets of modern references, different sets of SNPs, and different ancient samples (data not shown). Because the positional shifts were relatively minor and would not affect the conclusions drawn from the analyses, these were not followed up further. The data were converted to EIGENSTRAT format with the program convertf from the EIGENSOFT 7.2.0 package.[53] The results were plotted in R with ggplot2.[54]

Y chromosome analyses

In total, 113,217 haplogroup informative Y chromosome variants from regions that uniquely map to the Y chromosome[51,55,56] were called as haploid from the BAM file of the high-coverage TPM003 with --doHaploCall function in ANGSD.[49] Derived and ancestral allele as well as haplogroup annotations for each of the called variants were added via BEDTools 2.19.0[57] intersect option. Haplogroup R1a-YP578 assignment received the highest support of informative positions called in the derived state in TPM003. Further fine-level phylogenetic assignment was contextualized (Figure S8, Table S11) within the Y chromosome variation of Estonian high-coverage genomes[51] and the phylogenetic tree of Yfull YTree v.8.10.0.

LSAI sharing and individual connectedness inference

LSAI segments and kinship coefficients were estimated from merged PLINK files of EstBB samples, 503 Europeans from the 1000 GP, and 15 ancient Estonain genomes with IBIS[16] (identical by descent via identical by state) v.1.20.6 using –min_L 5 cM and –c 0.0005 kinship coefficient cut-offs—corresponding to minimum requirement of one shared segment of >5 cM length and total sharing of at least 0.1% of the genome ~6.6 cM—for most of the analyses (except for Table S4 comparing 2, 5, 7, and 10 cM thresholds) and –maxDist 0.1 and –mt 300 parameters. The total number of SNPs used varied between 244,643–254,326 MAF > 0.05 variants.

Although IBIS has the highest IBD inference accuracy for >7 cM segments,[16] we use >5 cM threshold in our diachronic inferences because our focus is on relationships at generational distances > 15 at which longer IBD block sharing expectations become relatively low,[58] particularly in combination with the loss of sensitivity to detect long IBD segments from imputed ancient DNA sequences, as shown by the fragmented nature of TPM003's self-sharing in Table S6. Because true IBD segments of this length are not expected to be common at these generational distances, we need to consider the detected segments as "long shared allele intervals" (LSAIs) rather than IBD segments sensu stricto. Because they are inferred from unphased data after removal of rare variants (which cannot be imputed with sufficient accuracy), the LSAIs are likely to include undetected recombination points and smaller IBD segments residing on different haplotypes. To control for potential effect of differences in sites with missing data on the LSAI inference, we analyzed all low-coverage (<0.3×) samples individually (in order to avoid cumulative loss of SNP numbers) against the EstBB and 1000 GP data by using the –setIndexEnd option in IBIS after filtering out, on individual basis, variants for which the low-coverage genome had missing data.

Further details on the LSAI inference parameter choice are given in the supplemental methods.

The probability of individual connectedness (PiC) score for individual x in group Z was estimated as the proportion of individuals from group Z with whom individual x shared IBD above the given threshold. In practice, we estimated the count of connected individuals from group Z from sorted IBIS .coef output files by using the linux "join" function to add group codes to individual identifiers and by using the "crosstab" function of datamash[59] to generate the table of counts, each of which we divided by the total number of individuals in group Z to obtain the individual connectedness proportions by groups (the PiC scores).

Simulations of LSAI sharing under different demographic models

To investigate the patterns of IBD sharing between contemporary and ancient samples expected under different demographic scenarios, we simulated eight different demographic scenarios, described in Table S9, by using msprime.[60] In all simulations, we used the discrete time Wright-Fisher model (model = "dtwf") to simulate generations 0 to 1,000 and then switched to the Hudson model (model = "hudson") as advised by msprime documentation for simulations with large sample size and multiple chromosomes. We used a recombination map obtained by concatenating two 1000 GP maps for chromosome 1 (GRCh37) separated by a region of 50 cM to increase analyzed sequence length. Mutation rate was set to 1.25×10^{-8}. In each simulation, we sampled 400 haplotypes (200 diploid samples) per time point per population (in the case with two populations simulated) at six time points: 0, 10, 20, 30, 50, and 100 generations ago. We filtered out positions falling into telomeric or centromeric regions of the chromosome 1 recombination map or in the junction between the two maps as well as positions with derived allele frequency less than 5% in the simulated dataset to match the filtering scheme applied on empirical data. The LSAI segments were detected via IBIS[16] with the same thresholds (at least one >5 cM shared segment and

kinship coefficient > 0.0005) as used in the analyses of the empirical data.

Unsupervised community extraction analyses

We used the list of individual pairs sharing at least one >5 cM LSAI segment and having a kinship coefficient > 0.0005 from the IBIS (.coef) results as an input (with three columns: id1, id2, kinship coefficient) for community extraction analyses. This list was passed to a custom R script that runs a hierarchical clustering method for "community detection," known as the Louvain algorithm,[17] that is implemented in the R library "igraph."[61] We introduced an additional step to quantify the significance of the extracted communities. Five nested cycles of the Louvain algorithm were run on each community passing the Wilcoxon rank-sum significance test implemented in the R library exactRankTests.[62]

In our pipeline, the igraph algorithm first detects all possible level-one communities and then each community undergoes a Wilcoxon rank-sum test that weighs the internal and external degrees of the community connections in order to quantify its significance. In cases of significantly (p value < 0.05) more internal than external connections, the communities are accepted and passed on to the analyses at the next level. All individuals from the communities that do not pass significance testing are excluded from further steps. Every next cycle of community extraction begins with modularity detection followed by testing of statistical significance before moving to the following cycle. We let this process continue up to the fourth level. A fifth cycle is internally implemented only for testing the statistical significance of the level-four communities. By the end of this process, a network of connections will include all those communities statistically supported at each level and a per community list of included individuals. At this point, using a series of custom scripts, we combined all the community levels in order to assign each individual to a community defined by a unique alphanumeric code, resuming a sample's complete path from one level to another. Based on the significance test results, each sample's last community level assignment can be confirmed (the sample maintains its position) or be changed (sample is reassigned to the previous statistically significant level). Finally, connectivity scores are estimated for individuals of the extracted communities.

PiC score was used for the outlier detection process in each extracted community. We screened by community the individuals for their PiC scores and identified as "outlier candidates" individuals below the lower whisker of the boxplot distribution of the PiC scores in the community they were assigned to. Each list of "outlier candidates" was tested against the overall distribution of the PiC scores in that same community via a custom R script for the significance. Communities with more than 25 individuals were tested with a Rosner's test, whereas communities with 25 or less members were tested with Dixon's test. Individuals with p values < 0.05 are marked as significant outliers and removed from further analyses. Eventually, 330 samples out of 4,852 were removed from the intensity score matrix as outliers, including 281/320 Slavic- or Baltic-speaking EstBB participants and 49/4,419 ethnic Estonians (Table S15). Community membership proportions were then plotted as pie charts in R with the ggplot2 package.

Phenotype prediction analyses

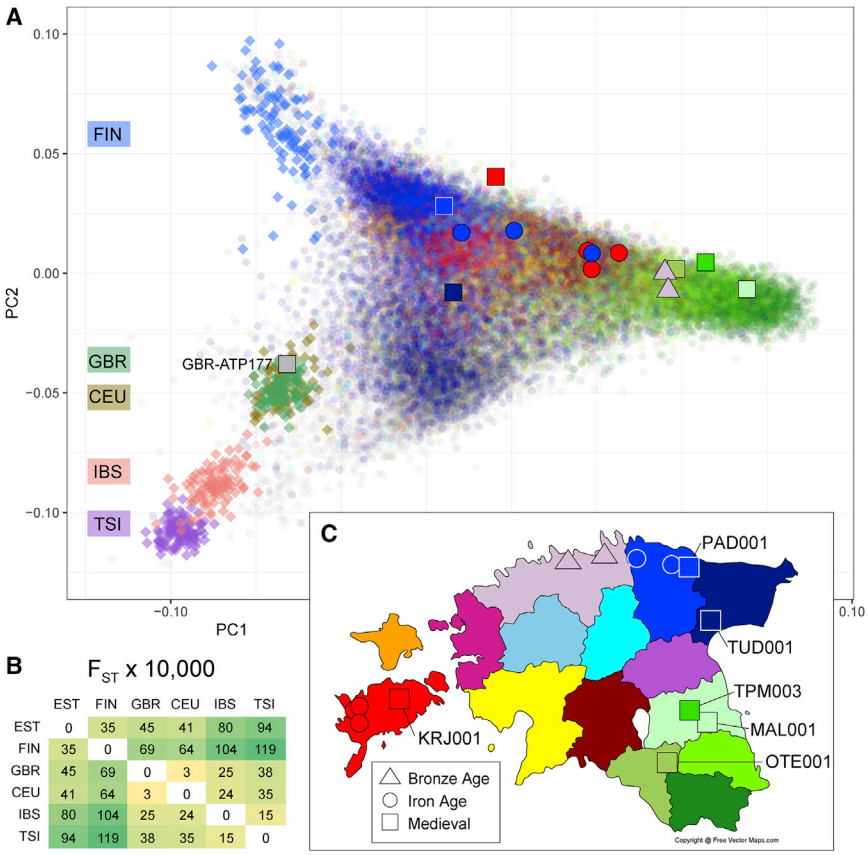
We used vcftools[63] (--snp option) to extract the genotype information at 104 phenotype informative markers already analyzed,[64] after excluding nine SNPs absent in our Estonian reference panel, from the high-coverage TPM003 and for its down-sampled (0.1× and 0.3×) and imputed copies. We then filtered the genotypes to keep only variants with GP ≥ 0.99 and recorded them as the number of effective alleles by using PLINK (--recode A --recode-allele option). For the HIrisPlex-S set for the pigmentation prediction,[65] we uploaded the genotype data to the HIrisPlex-S webtool after reformatting by using the "merge" function in R to combine information from all informative SNPs. We interpreted the results of the webtool according to its manual to obtain the pigmentation prediction (Table S13).

Further details of the phenotype prediction concordance estimation and analyses of the *SLC24A5* region are given in the supplemental methods.

Results

While IBD-based methods can offer high-resolution insights into the recent phases of our demographic history, the accuracy and robustness of shared IBD inferences—or the related signal we explore here, LSAI—from low-coverage ancient genomes has not been determined yet. To address this, we sequenced the genome of a 15th century male individual (TPM003) from Tartu Püha Maarja (St. Mary) parish cemetery (Estonia) to an average depth of 23× (Table S1). We determined the genotype calls of the high-coverage genome of TPM003 directly and then compared the results against genotype calls from five down-sampled copies of 0.1× coverage, each of which we imputed by using a panel of 2,076 Estonian high-coverage sequences.[51] We estimated the average proportion of matching heterozygote calls between the imputed and high-coverage data as the primary estimator of imputation accuracy at 98.6% for common (MAF > 0.05) variant sites and noticed a notable accuracy drop to <95% and <80% in variants with MAF < 0.05 and MAF < 0.01, respectively (Table S2). In further analyses, we used only variants with MAF > 0.05.

We next analyzed the imputed low-coverage ancient samples from Estonia and one medieval sample from the UK together in context of genotype data from 69,218 individuals from the EstBB and 503 Europeans from the 1000 GP[40] by using FlashPCA2[52] and smartpca[53] (Figure 2 and Figure S1). We observed a clear distinction of Estonians from other European populations, including Finns (Figure 2A). By the fixation index (F_{ST}) statistic, Estonians are more differentiated from neighboring Finns than, for example, 1000 GP Italians from Tuscany are from the Iberians (Figure 2B). We found that all 14 imputed Bronze Age, Iron Age, and medieval samples from Estonia cluster together with present-day Estonians approximately within the same broad geographic regions of Estonia in which they were buried, although the resolution afforded by these analyses did not allow for finer county-level assignments (Figure 2D). Similarly, we found that the medieval British sample that we imputed together with five Estonian medieval genomes maps close to the GBR cohort (Figures 2A



	EST	FIN	GBR	CEU	IBS	TSI
EST	0	35	45	41	80	94
FIN	35	0	69	64	104	119
GBR	45	69	0	3	25	38
CEU	41	64	3	0	24	35
IBS	80	104	25	24	0	15
TSI	94	119	38	35	15	0

Figure 2. Principal-component analysis of ancient and modern Estonian genomes
(A) Principal-component map of 69,218 modern Estonian Biobank (EstBB) individuals, 14 imputed Estonian genomes, and one imputed British (GBR-PSN177) genome shown in context of 503 Europeans from the 1000 GP. The proportion of total variance explained: PC1, 0.00094; PC2, 0.00058. The coordinates of the ancient genomes were calculated directly without use of projection.
(B) F_{ST} estimates among modern populations where EST refers to Estonians from the EstBB and other abbreviations to those of the 1000 GP populations.
(C) Geographic locations of modern individual birth and ancient individual burial places in Estonia are shown in the map with colors corresponding to those used in the principal-component map. Further details of ancient genomes are provided in Table S1.

and 2C). We confirmed the robustness of the placement in PCA of ancient imputed genomes directly without the use of projection by comparative analyses of high-coverage, imputed, and haploid-called and projected data (Figure S2). Notably, the down-sampled replica of TPM003, imputed from 0.1× coverage, mapped next to the closest neighbors from the EstBB in the PCA constructed with the high-coverage sample without imputation, suggesting that high accuracy (with less variance than from projections of haploid-called genotype data) of individual ancestry mapping is possible from imputed data at this coverage (Figure S2).

To explore regional LSAI sharing patterns, we used IBIS[16] to extract pairs of individuals who share long unphased (>5 cM) LSAI segments and estimated kinship coefficient > 0.0005. We introduce PiC, the probability of individual connectedness, as a simple measure to explore patterns of LSAI sharing within and among populations by user-defined segment length (L) and kinship coefficient k (as a measure of total genome-wide IBD sharing) thresholds. We first compared the outgroup-f3 statistic as a measure of drift sharing against PiC among Estonians, Baltic and Slavic speakers from EstBB, and Finns from 1000 GP and observed that PiC offers high resolution in distinguishing local differences (Figure S3). There is a notable decline of within-region connectedness across year of birth cohorts (Figure S4), which most likely reflects higher mobility within Estonia in the last few generations. Consistent with the lack of major geographic barriers and proximity, Estonians share most drift with Baltic-speaking Latvians, while by the PiC statistic, Finns are the most closely connected group to Estonians. The differentiation of Estonians and Finns by drift-sensitive statistics, such as f3, can be explained by the founder effects in the Finnish demographic history, a finding that is consistent with the higher genetic F_{ST} differentiation between Estonians and Finns than among Tuscan and Iberian genomes (Figure 2B); the higher level of Estonian connectedness with Finns than with Baltic-speaking neighbors by PiC requires, however, further scrutiny with regards to the time depth of these connections. Analysis of ancient genomes can provide answers about whether the LSAI segment sharing reflects recent gene flow or some other aspects of shared demographic history.

We assessed IBD sharing between ancient and modern genomes from Estonia and found that the proportion of individuals from the EstBB with whom ancient genomes share LSAI segments increased from Bronze to Iron Age and medieval periods (Table S3). We observed significantly higher PiC scores between EstBB and medieval (12^{th}–16^{th} centuries) than EstBB and Iron ($p = 0.02$, two-tailed t test) or Bronze Age ($p = 0.017$) samples when we used LSAI length > 5 cM threshold. These results stand in contrast to the lack of clear differences between the diachronic samples in their >2 cM LSAI sharing (Table S4). Further, we observe that the sharing of >5 cM and >7 cM LSAI is at comparable levels for modern-modern and modern-medieval pairs of samples, while >10 cM segments can be detected more abundantly in modern-modern than modern-ancient pairs, most likely because of the excess of distant genealogical relationships among the modern samples that would be absent in modern-ancient pairs.

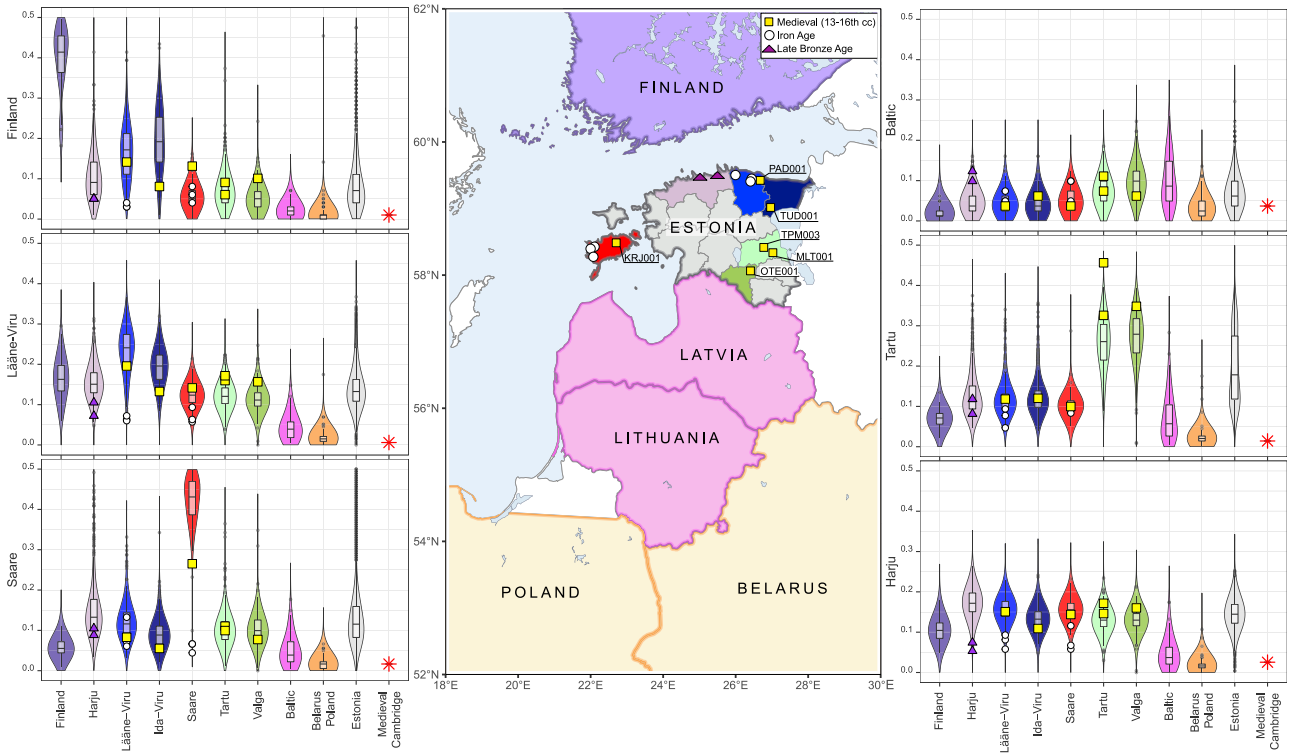


Figure 3. Genetic connectedness with six target populations of the circum-Baltic region and Estonian Bronze Age, Iron Age, and medieval genomes

Each individual violin plot shows the distribution of PiC scores, which reflect individual probabilities of >5 cM LSAI sharing and kinship coefficient estimate higher than 0.0005. Individuals from the modern target population are shown on the y axis. Distributions of the present-day populations are shown by the colors according to the map. Present-day genomes include 99 Finnish individuals from the 1000 GP data; 4,739 EstBB Estonians born before 1940, including 1,880 from Harju, Saare, Viru, Tartu, and Valga counties, and 320 EstBB Latvians, Lithuanians, Belarusian, and Polish individuals born outside Estonia. All ancient samples, shown with squares (medieval), circles (Iron Age), and triangles (Bronze Age), have been imputed, including one medieval 0.1× coverage British genome PSN177 as a control, with 2,092 Estonian high-coverage sequences as a reference panel.

To explore further regional details of LSAI sharing patterns between Estonians and their geographic neighbors in light of evidence from ancient imputed genomes, we focused on $PiC_{L > 5\ cM,\ k > 0.0005}$ scores in a subset of Estonians of the EstBB born before 1940 for whom county- and parish-level information of birthplace was available (Figure 3, Figure S9). Under realistic scenarios of human population densities and dispersal rates, virtually all pairwise shared IBD blocks longer than 4 cM are expected to coalesce to common ancestors within the last 100 generations[66] or approximately 3,000 years. Consistent with this prediction, we observe marginally low levels (1%–2%) of >5 cM LSAI sharing between Estonians and other East European populations (Poles, Belarusians, Lithuanians) (Figure 3) with whom they share Steppe ancestry[31] through Late Neolithic dispersals from a common Corded Ware culture source (Figure 1C).

Estonian Iron and Bronze Age individuals sampled at a 2,400–2,800 years time depth show >10× higher connectivity with modern Finnic- and Baltic-speaking populations than with West Europeans (Table S3), while their PiC scores with present-day Estonians are lower than with Baltic-speaking Latvians and Lithuanians (Table S3, Figure 3). These observations are in line with common ancestry sharing in a broader area of Corded Ware culture before the arrival of Finnic speakers: consistent with this model (Figure 1C), Belarusians and Poles show more LSAI sharing with the Bronze Age than present-day individuals from Estonian (Figure 3). We observe no significant excess of PiC scores between EstBB Estonians and Iron or Bronze Age genomes sampled from the same Estonian counties (Table S10, Figure 3). Neither do we see higher regional affinity between North Estonian Bronze and Iron Age samples in relation to Iron Age samples from Saaremaa. Overall, these results suggest that the present-day county-level LSAI sharing patterns were not yet fixed in the Iron Age.

In contrast to individuals sampled from earlier periods, six Estonian medieval genomes from the 13th–16th centuries share significantly more >5 cM LSAIs with present-day individuals born in the same county in Estonia (Figure 3, Figure S9). Furthermore, all ancient genomes from Estonia studied here show high affinity not only to present-day Estonians but also to present-day Finns at levels up to an order of magnitude higher than to Swedes (Table S10), including Late Bronze Age (average 5%) and Iron Age (average 5%, range 3%–8%) genomes. Estonians share more Finnish LSAIs than 82 EstBB Latvians and Lithuanians (average 2%, range 0%–7%) or a medieval

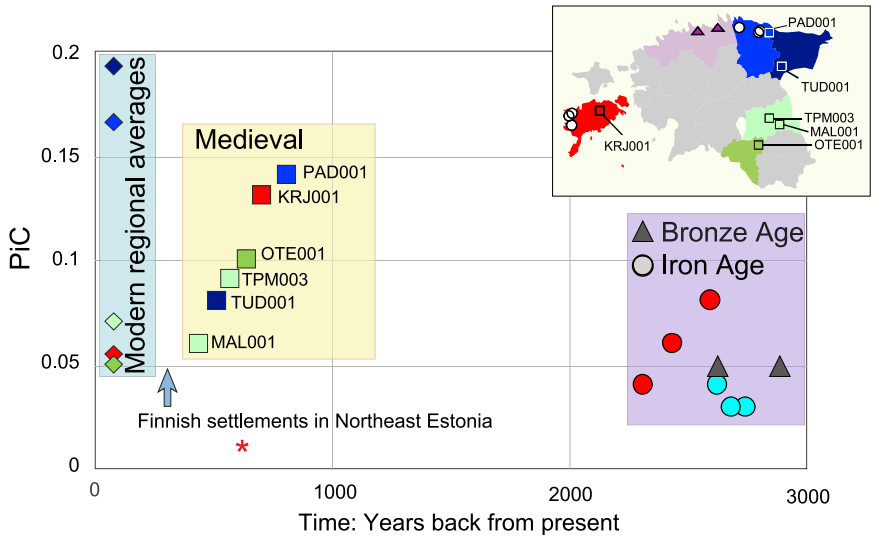


Figure 4. Temporal changes of individual connectedness with Finns in Estonian genomes from the Bronze Age to present day

Medieval, Bronze Age, and Iron Age individuals are each represented by a square, triangle, or a circle, respectively. Individual connectedness with Finns (PiC_{FIN}) of individuals born before 1940 is shown by diamonds for selected counties from which the ancient genomes were sampled. The birthplace of present-day individuals and the burial places of the medieval genomes are indicated by color according to the shown map. PiC, proportion of 1000 GP Finnish individuals with whom modern and ancient Estonians share at least one LSAI segment > 5 cM and kinship coefficient > 0.0005; *, medieval British low-coverage (0.1×) genome PSN177 imputed together with medieval Estonians with reference panel of 2,092 Estonian high-coverage whole-genome sequences.

low-coverage genome from Cambridgeshire, UK imputed with Estonian medieval samples (Figure 4). Medieval Estonians, however, share significantly more >5 cM LSAIs with Finns (10.1% average, Table S3) than Iron Age (average 4.7%, p = 0.006, two-tailed t test) or modern Estonians (8.7% on average), suggesting that recent (17^{th}–18^{th} centuries) localized migration events cannot explain the excess of Finnish LSAI sharing that we observe across Estonia today. Instead, these findings point to a migration event across the Gulf of Finland earlier than the 13^{th} century as being responsible for the observed patterns.

We observe higher consistency in regional LSAI sharing patterns between the high-coverage genome and its down-sampled replicas (Table S8, for further details see supplemental methods). To summarize the compound effect of imputation errors on LSAI-based ancestry mapping of the ancient samples, we applied the uniform manifold approximation and projection (UMAP) dimension reduction method on the regional (county-based) PiC scores (Figure 5) and observed that the imputed TPM003i0.1× mapped closely to its high-coverage version that had not been imputed. This, along with regional clustering of other medieval genomes among EstBB individuals from the same geographic context suggests that (1) LSAI inference through imputation from ancient low-coverage genomes can be achieved sufficiently accurately for addressing questions about regional ancestry; (2) >5 cM LSAI segments persist and remain regionally informative for at least 800 years, over which time the regional genetic identities in Estonia have remained relatively distinct from one another; and (3) considering the fact that local Iron and Bronze Age populations do not show region-specific affinity to present-day local communities—although most likely being genetically ancestral to these in a broader geographic sense—these regional LSAI sharing patterns that unite medieval and present-day Estonians were most likely created between the Iron Age and the 12^{th}/13^{th} centuries AD whence our earliest medieval samples derive.

The geographic patterns of connectedness we estimated from PiC scores of 15 present-day counties of Estonia rely on administrative divisions that may have not been meaningful in the past. To further test the robustness of the inference of geographic patterns in our data, we used an unsupervised modularity optimization technique, called the Louvain method,[17] that clusters individuals into modular units (communities) by their LSAI connectivity among individuals without the use of any geographic or other sample pooling criteria. We extracted communities by using a nested application of the Louvain algorithm, allowing each detected community to undergo a further cycle of community identification on the basis of significant excess of internal as opposed to external connections. We applied the Louvain method on the IBIS results for 4,739 EstBB donors born before 1940, 14 ancient genomes from Estonia, and 99 Finns from the 1000 GP. The Louvain method revealed four first-order and 20 second-order communities, which roughly corresponded to the main geographic regions of Estonia (Figure 6). Notably, all Finns of the 1000 GP data clustered together in one of the 3^{rd}-level sub-clusters, I7b, of a 2^{nd}-level cluster, I7, that has predominantly Northeast Estonian provenance (Figure 6, Table S15). The Louvain method places all six medieval genomes from Estonia into communities containing modern genomes from the same geographic region as their burial place while lumping all eight Iron and Bronze Age genomes, regardless of their geography, to community I4 (Figure 6, Table S15). The I4 community contains a small number of modern counterparts and is characterized by low connectedness both internally and to other communities, which is uncharacteristic of modern and medieval genomes (Table S15).

We next ran simulations with msprime[60] in order to better understand the observed patterns of extensive LSAI

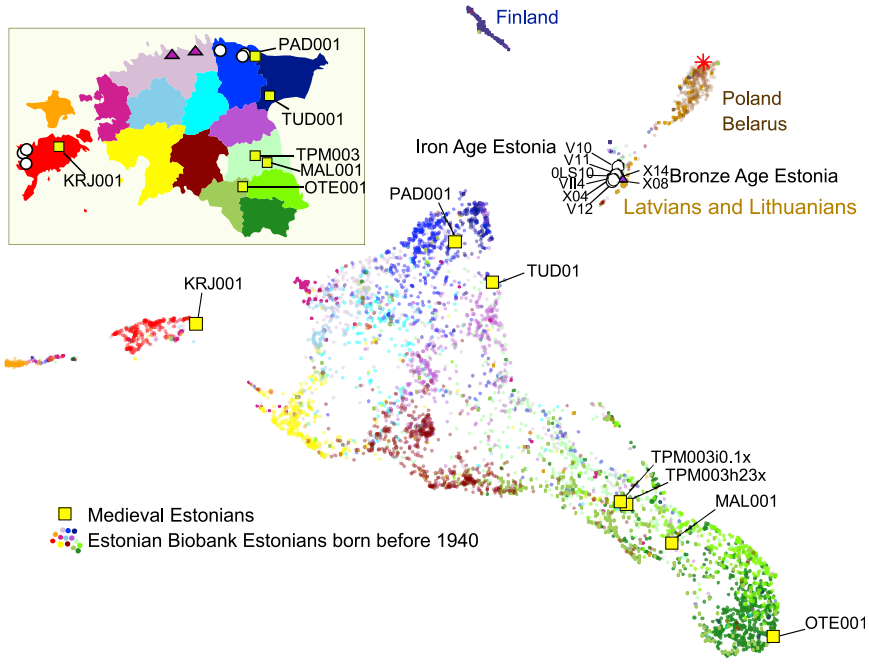


Figure 5. UMAP plot of LSAI sharing among medieval, Bronze Age, and Iron Age genomes from Estonia; EstBB Estonians born before 1940; 1000 GP Finns; and EstBB Latvian, Lithuanian, Polish, and Belarusian samples

UMAP analyses were performed on 18 PiC score vectors estimated with L > 5 cM and k > 0.0005 thresholds for 15 Estonian counties, FIN, Baltic, and Slavic speaking groups. County map of Estonia is shown with ancient sample locations. Genotypes of all ancient genomes shown, except for TPM003 h23x, were imputed. Six medieval Estonian genomes were imputed together with a medieval British low-coverage (0.1×) genome PSN177 shown with red asterisk (*).

sharing between ancient and modern Estonian genomes and how these are affected by demographic history (Figures S5 and S6). The simulation models were inspired by effective population size (Ne) trajectories obtained by applying IBDNe[67] to modern high-coverage Estonian genomes.[38] We show (Table S9) that Ne and its changes over time significantly affect the pattern of IBD sharing between individuals. First, unsurprisingly, the results of the simulations show that the fraction of the population that an average individual is connected to is inversely and linearly dependent on Ne (compare Figures S6A and S6C versus Figures S6F and S6H), resulting in little expected connectedness in large populations. Second, under a population model with a recent exponential growth, modern individuals can have a higher LSAI sharing with ancient individuals sampled from periods of small population size preceding the growth compared to IBD sharing with present-day individuals; the specific pattern is dependent on the duration of the growth period and the growth rate (Figure S6). Third, under scenarios realistic for Estonian subpopulations, we find that present-day individuals are expected to have similarly high levels of LSAI sharing with their contemporaries and with ancient individuals from up to 30 generations (~900 years) ago at maximum, and there is a notable drop deeper in time (Figures S5A and S5C, Figures S5B and S5D, and Figures S5E and S5G). This means that our simulations do not support a model by which the high connectedness between Finns and Estonians could derive from Iron Age migrations circa 100 generations ago (Figure 1A). Finally, under a simplistic model of a clean population split with no subsequent gene flow, present day individuals from one of the populations are expected to show an increasing level of IBD sharing with individuals from the other population as we sample from time points successively closer to the split time (Figure S5F and S5H, blue boxes). The latter observation explains why present-day Finns can have higher IBD sharing with medieval rather than contemporary Estonians from certain Estonian regions.

The high-coverage genome of TPM003 allowed us to examine his Y chromosome at high resolution in context of a large reference set of 1,160 high-coverage sequences from Estonia.[51] Consistent with the autosomal LSAI sharing results, we detect a signal of regional clustering of TPM003 together with lineages from Southeast Estonia in a newly defined R1a1c-B2153 clade (Figure S8, Table S11), which is nested within a broader set of Y chromosomes in a clade R1a1c-YP578. According to YFull tree, this clade has been estimated to have a coalescent date of 2,100 (confidence interval [CI] 95% 2,300–1,800) years and geographic distribution mainly in present-day Russia and Finland. Although R1a1c-YP578 is widely spread across Estonia (3.1% on average), the newly defined R1a1c-B2153 has, according to our knowledge, not yet been found outside of Estonia. Among 1,160 Estonian Y chromosomes, it is only found in six individuals from Tartu and Põlva counties, including a grandfather-father-son trio from Tartumaa. Unsurprisingly, considering the generational distance between the ancient and modern genomes, we find no evidence of triangular autosomal IBD transmission of the medieval TPM003 shared segments from the modern grandparent to his grandchild.

Because imputation of genotypes at loci that have been targets of selection can be problematic[68] and we had not filtered out such variants from our analyses, we assessed the accuracy of imputation at 104 functionally informative positions widely used for phenotype inference, including those affected by recent selection (Table S12). We observed high (>0.98) match rate between imputed and high-coverage genome for genotype calls at 90 variants that were sufficiently well covered in the high-coverage TPM003 genome (Table S13).

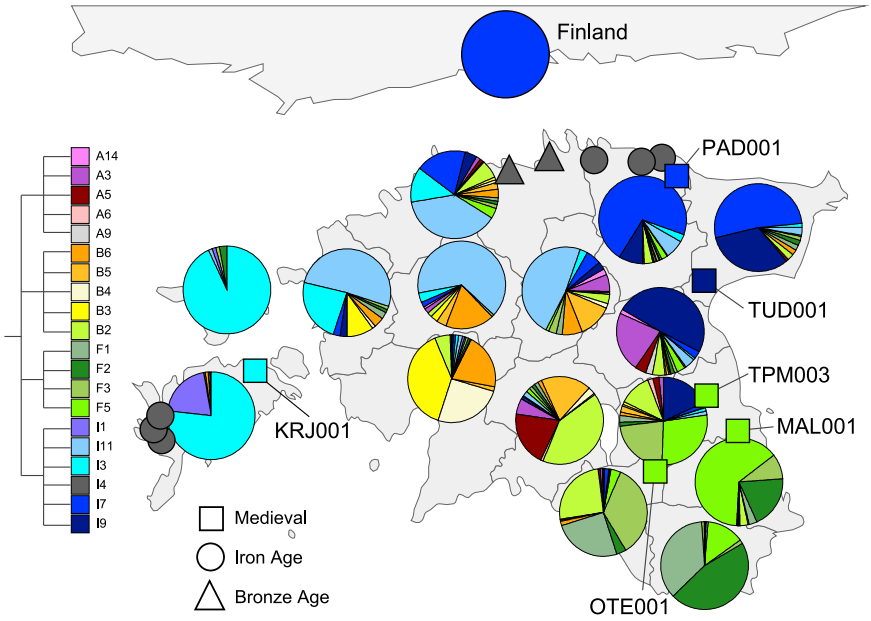


Figure 6. Two-level community structure inferred from 4,739 Estonian Biobank individuals born before 1940, Finns of the 1000 GP, and 14 ancient Estonians
Unsupervised community extraction method (Louvain method[17]) was applied on 5 cM LSAI-sharing signals estimated with IBIS. Four communities with more than ten members (A, B, F, and I) are detected at the first level of extraction. One of the communities, I, which is most widespread in North and West Estonia, was further divided into six significant subcommunities (I1, I3, I4, I7, I9, and I11), which have more specific regional distributions. One of these, I7 is more common in Northeast Estonia where it has six major sub-communities (Table S15). The I7 community also captures all 99 Finns of the 1000 GP. Pie charts on the map show the sub-community membership proportions in 15 Estonian counties, the FIN subset of the 1000 GP data (without regional detail of administrative units within Finland), and 14 ancient genomes from Estonia.

Interestingly, we found TPM003 to be heterozygous at rs1426654 (A/G) in the *SLC24A5* that has been identified among 22 strongest signals of selection in human genome.[69] rs1426654 is a variant that explains a major part of skin pigmentation differences between Africans and Europeans and differences among South Asians.[70–72] The derived A allele at this variant, associated with lighter pigmentation, has been shown to have been introduced to Europe by Neolithic farmers followed by its virtual fixation in most European populations today.[73] The highest frequency of the ancestral G allele in the 1000 GP Europeans appears to be in Finns (2%).

Because genotype imputation at variants with low MAF is reduced and potentially problematic in regions of the genome targeted by natural selection,[68] we further assessed the accuracy of our LSAI inference from imputed data by comparing TPM003's LSAI sharing, with a >2 cM threshold, around the rs1426654 variant in the local Estonian reference panel that was used in imputation against LSAI matches for TPM003 in this locus in the Haplotype Reference Consortium (HRC) and the 1000 GP panels not used in our imputation. We found that both directly called and imputed copies of TPM003 share IBD segments, both for the A and G allele, with Estonian and Finnish samples from different haplotype panels (Figure S7, Table S14). Among segments longer than 5 cM, there are both G- and A-allele-carrying haplotypes shared between TPM003 and four Estonians (three with the A allele and one with the G allele) and one Finn (with the G allele) (Figure S7A). On average, G-allele-carrying haplotypes are significantly longer (3.44 cM versus 2.42 cM, one-tail t test: 6.63×10^{-9}), suggesting more recent common ancestry in Finns and Estonians of the ancestral than the derived allele, which has been highlighted as one of the strongest targets of positive selection in human populations.[69] The core 200 kb G haplotype observed in TPM003 is distinct from Asian and African G haplotypes and observed in the given sample only among Estonians and Finns (with the exception of a single Japanese [JPT] individual from 1000 GP) (Figure S7B).

Discussion

We have shown that shotgun sequencing of ancient DNA at low (0.1–1×) coverage enables sufficiently accurate genotype data imputation for ancestry and IBD/LSAI-based community structure analyses. Our estimated imputation accuracy of 0.99 (Table S1) for heterozygote calls of common variants from a medieval Estonian genome is higher than the 0.93 estimate we previously obtained by applying the same approach[10] on a high-coverage Neolithic genome[8] and higher than comparable estimates of accuracy of <0.90 for other approaches.[6,74] The increased accuracy can potentially be explained by (1) the temporal (and genetic ancestry) proximity of our medieval genome compared to the Neolithic sample considered in Hui et al.[10] and (2) the use of a large, ethnically/regionally matched reference panel. Although our analyses at phenotype-informative variants, including those highlighted previously as selection targets, did not reveal a notable drop in imputation accuracy, we caution against generalizations to cases without ethnically/regionally matched large reference panels and longer time gaps between the imputed sample and the imputation panel and note that these results are based on a limited number of observations of a diverse set of 90 different variants taken between a single high-coverage genome and its down-sampled replicas.

Downstream PCA and LSAI analyses showed sufficiently high precision for fine-scale mapping of the genetic

ancestry of the imputed samples. Our analyses showed relatively lower accuracy of IBD1 and IBD2 recovery from imputed low-coverage (0.1×) genomes, suggesting that detection of sample identity, as well as twins and 1st-degree relatedness from imputed low-coverage genomes via IBD/LSAI approach, can be challenging; however, at 0.3× coverage, we were able to correctly recover 92.6% of IBD2 and 96.5% of total IBD segments of the down-sampled ancient genome (Table S6). Furthermore, other methods such as READ[75] or GRUPS[76] offer accurate estimation of close relatives from low (>0.05×) coverage data.

Kinship coefficients are a measure of the proportion of genome-wide IBD in a pair of individuals. The abundance of long IBD segments can be a robust indicator of close relatedness in a large unstructured population and is therefore widely used by direct-to-consumer genetic testing for inferring matches in genealogical relationships (up to 5th cousins). However, our simulations (Figures S5 and S6) show that IBD sharing patterns are strongly influenced by effective population size history. The kinship coefficients estimated here between modern and medieval samples are clearly not interpretable in terms of meaningful genealogical relationships given that the pairs are separated by more than 15 generations in time. Hence, the signals of elevated LSAI sharing (comparable in their intensity to the levels of 4th- to 5th-cousin relationships in large populations) between present-day individuals and those sampled 10–20 generations ago can be best explained, in line with our simulation results, by relatively low historic Ne and recent exponential growth in Estonia. Additionally, a large fraction of the segments we detect may not correspond to shared haplotypes because of their unphased nature and their length,[39] and as such, represent a series of very short segments that coalesce, on average, longer ago than a true IBD segment of the same length would. Consistent with this, among the triangular cases, where a medieval genome shares LSAI with two modern individuals who are themselves closely related (grandparent-parent-offspring sequence), we observe no excess of LSAI sharing in the grandparent (Figure S8), which would be expected under genealogical relationships. Thus, it is more likely that most cases of diachronic LSAI sharing that we describe are explainable by cumulative long-term maintenance of community-specific chunks of IBD through marriages involving distant (cryptic) relatedness within the same parish- or county-level community.

We suggest that diachronic LSAI sharing patterns can be informative for resolving complex demographic scenarios involving recent population splits and subsequent gene flow. Most shared IBD blocks longer than 4 cM are expected to be less than 1,500 years old,[2] and virtually all IBD blocks of this length are expected to derive from the last 3,000 years.[66] The large-scale >5 cM LSAI sharing between Estonians and Finns would thus be expected to reflect primarily historical gene flow across the Gulf of Finland (Figure 1F) while not necessarily standing in conflict with contributions from earlier migration events predicted on linguistic grounds (Figure 1E) or synthesis of archaeological and genetic evidence (Figure 1D).[19]

The high levels of LSAI sharing with Finns that we observe in present-day Northeast Estonians could, at least partly, be explained[38] (Figure 1F) by historically attested Finnish settlements in Northeast Estonia in the 17th–18th centuries. However, our ancient DNA evidence (Figures 3 and 4) from the 12th-16th centuries points to deeper time depth for this relationship across Estonia. According to the current synthesis of genetic and archaeological evidence, the earliest migration event that could account for genetic ancestry sharing and unique connectedness among Finnic-speaking Finns and Estonians dates back to the Pre-Roman Iron Age (Figure 1D [19]). However, the Nganasan-related autosomal component that appears in the circum-Baltic region in this time period as a signature of possibly the first arrival of Finno-Ugric speakers is likely to have reached Fennoscandia and Estonia by different routes and is relatively minor (3%–5% of total autosomal ancestry).[19,32,20] Yet, our analyses of ancient genomes through the transect of time show that the levels of LSAI sharing with present-day Finns have been higher among Estonian genomes than those observed in present-day Latvians and Lithuanians (Table S3) since not only the Iron Age but also the Bronze Age (Figures 3 and 4), suggesting they have been generated *in situ* in Estonia for a long period of time rather than being introduced to Estonia from external sources recently.

The minor Nganasan-related component in the Pre-Roman Iron Age migrations (Figure 1D) could explain the specific G-allele-carrying haplotype distribution of *SLC24A5* among Finns and Estonians (Figure S7), but the genome-wide sharing patterns, in which individuals from the 12th–14th centuries AD show the highest connectedness to present-day Finnish genomes (Figure 4), are arguing against the Pre-Roman Iron Age time depth for the main connectedness signal we observe. Furthermore, the results of our community extraction analyses (Figure 6) suggest that the patterns of region-specific connectedness within Estonia postdate our Iron Age samples and that all 99 Finnish samples we explored were assigned to a single community primarily composed of North Estonians. Notably, in simulations, this diachronic pattern of extensive sharing between past and present populations is consistent not only with the outcome of a population split model (Figure S6F) but also observable under a range of panmictic cases that consider realistic demographic scenarios of population history in Estonia (Figure S6E). In sum, these results suggest that informative LSAI signals can persist in structured populations at least for dozens of generations and that the high level of connectedness between North Estonian and Finnish genomes is older than our earliest medieval and younger than our latest Pre-Roman Iron Age samples.

The high level of Finnish LSAI sharing observed in individuals who lived in Estonia during the 12th–14th centuries

AD represents the first direct evidence that a significant proportion of these relationships date back to the time before the expansion of the Finnish population, the Finnish founder event,[29] i.e., to the time when the total population size of Finland is estimated to have been very small. Because we observe nearly identical and highly correlated ($r > 0.999$) levels of LSAI sharing between Estonian counties and Finnish samples from two independent datasets (Table S10), we consider our results to be robust and representative for the Finnish population in general. However, considering the existence of significant population substructure in Finland,[36–38] further research would be required for determining regional and temporary details of the connectedness patterns revealed here within the context of temporary changes of ancestry and substructure of the Finnish population.

In sum, the results of our analyses on genetic data are consistent with the linguistic model (Figure 1E) that ascribes the language affinities and innovations shared between the Finnish language and North Estonian dialects to a migration event from North Estonia to Finland in the end of the first millennium. Because LSAIs are expected to decay in time because of recombination and admixture, the fact that present-day Finns still show genetic connectedness with medieval and modern North Estonians at levels comparable to internal connectedness in Estonia suggests that these uniquely shared long allele intervals are abundantly present across the genome as a feature that characterizes a major part of Finnish genetic ancestry. However, more precise quantification of the impact of the migration event and its timing would require ancient DNA evidence from Finland before and after the event as well as modeling of Finnish effective population size history in context of its local regional diversity and admixture sources.

Data and code availability

The community extraction analysis scripts generated during this study are available at https://github.com/SABiagini/Louvain. The ancient genomic data generated during this study are available at https://www.ebi.ac.uk/ena/browser/view/PRJEB46155 (accession code ENA: PRJEB46155) and the data depository of the EBC (https://evolbio.ut.ee/). The Estonian Biobank (EstBB) data used in this study are available under restricted access. The procedure of applying for access to the data can be found under the following link: https://genomics.ut.ee/en/biobank.ee/data-access.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.07.012.

Acknowledgments

We would like to thank the University of Tartu Development Fund for support to the Collegium for Transdisciplinary Studies in Archaeology, Genetics, and Linguistics. Analyses were carried out with the facilities of the High-Performance Computing Center of the University of Tartu. This work was funded by the Estonian Research Council grants PRG243 (M. Metspalu, Lauri Saag, C.L.S., Lehti Saag), PRG1027 (K.T.), PRG1071 (S.R.), and PRG29 (M. Malve); KU Leuven startup grant STG/18/021 (T.K.); KU Leuven BOF-C24 grant ZKD6488 C24M/19/075 (T.K. and S.A.B.); and The Wellcome Trust award no. 2000368/Z/15/Z (T.K., R.H., C.L.S.). D.N.S. and A.L.W. were supported by NIH grant R35 GM133805. E.D'A. was supported by Sapienza University of Rome fellowship "borsa di studio per attività di perfezionamento all'estero 2017," C.L.S. was supported by European Regional Development Fund 2014–2020.4.01.16–0030, and K.T. was supported by UT Institute of Genomics grant PP1GI19936. This research has been conducted with the UK Biobank resource under application numbers 54698 and 19947.

Declaration of interests

A.L.W. is a paid consultant for 23andMe and the owner of HAPI-DNA LLC. All other authors declare no competing interests.

Received: March 29, 2021
Accepted: July 23, 2021
Published: August 18, 2021

References

1. Lazaridis, I. (2018). The evolutionary history of human populations in Europe. Curr. Opin. Genet. Dev. *53*, 21–27.
2. Ralph, P., and Coop, G. (2013). The geography of recent genetic ancestry across Europe. PLoS Biol. *11*, e1001555.
3. Ferrando-Bernal, M., Morcillo-Suarez, C., de-Dios, T., Gelabert, P., Civit, S., Diaz-Carvajal, A., Ollich-Castanyer, I., Allentoft, M.E., Valverde, S., and Lalueza-Fox, C. (2020). Mapping coancestry connections between the genome of a Medieval individual and modern Europeans. Sci. Rep. *10*, 6843.
4. Browning, B.L., and Browning, S.R. (2016). Genotype Imputation with Millions of Reference Samples. Am. J. Hum. Genet. *98*, 116–126.
5. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. *103*, 338–348.
6. Rubinacci, S., Ribeiro, D.M., Hofmeister, R., and Delaneau, O. (2020). Efficient phasing and imputation of low-coverage 1 sequencing data using large reference panels. Nat. Genet. *53*, 120–126.
7. Cassidy, L.M., Maoldúin, R.O., Kador, T., Lynch, A., Jones, C., Woodman, P.C., Murphy, E., Ramsey, G., Dowd, M., Noonan, A., et al. (2020). A dynastic elite in monumental Neolithic society. Nature *582*, 384–388.
8. Gamba, C., Jones, E.R., Teasdale, M.D., McLaughlin, R.L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. Nat. Commun. *5*, 5257.
9. Martiniano, R., Cassidy, L.M., Ó'Maoldúin, R., McLaughlin, R., Silva, N.M., Manco, L., Fidalgo, D., Pereira, T., Coelho, M.J., Serra, M., et al. (2017). The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. PLoS Genet. *13*, e1006852.

10. Hui, R.Y., D'Atanasio, E., Cassidy, L.M., Scheib, C.L., and Kivisild, T. (2020). Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. Sci. Rep. *10*, 18542.
11. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.
12. Naseri, A., Liu, X., Tang, K., Zhang, S., and Zhi, D. (2019). RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. Genome Biol. *20*, 143.
13. Shemirani, R., Belbin, G.M., Avery, C.L., Kenny, E.E., Gignoux, C.R., and Ambite, J.L. (2019). Rapid detection of identity-by-descent tracts for mega-scale datasets. Nat. Commun. *12*, 3546.
14. Zhou, Y., Browning, S.R., and Browning, B.L. (2020). A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. Am. J. Hum. Genet. *106*, 426–437.
15. Dimitromanolakis, A., Paterson, A.D., and Sun, L. (2019). Fast and Accurate Shared Segment Detection and Relatedness Estimation in Un-phased Genetic Data via TRUFFLE. Am. J. Hum. Genet. *105*, 78–88.
16. Seidman, D.N., Shenoy, S.A., Kim, M., Babu, R., Woods, I.G., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., and Williams, A.L. (2020). Rapid, Phase-free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification. Am. J. Hum. Genet. *106*, 453–466.
17. Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks (J Stat Mech-Theory E.).
18. Saada, J.N., Kalantzis, G., Shyr, D., Robinson, M., Gusev, A., and Palamara, P. (2020). Identity-by-descent detection across 487,409 British samples reveals fine-scale population structure, evolutionary history, and trait associations. Eur. J. Hum. Genet. *28*, 2–3.
19. Saag, L., Laneman, M., Varul, L., Malve, M., Valk, H., Razzak, M.A., Shirobokov, I.G., Khartanovich, V.I., Mikhaylova, E.R., Kushniarevich, A., et al. (2019). The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. Curr. Biol. *29*, 1701–1711.e16.
20. Tambets, K., Yunusbayev, B., Hudjashov, G., Ilumäe, A.M., Rootsi, S., Honkola, T., Vesakoski, O., Atkinson, Q., Skoglund, P., Kushniarevich, A., et al. (2018). Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. Genome Biol. *19*, 139.
21. Lang, V. (2020). Homo Fennicus. (Helsinki: Suomalaisen kirjallisuuden seura).
22. Honkola, T., Vesakoski, O., Korhonen, K., Lehtinen, J., Syrjänen, K., and Wahlberg, N. (2013). Cultural and climatic changes shape the evolutionary history of the Uralic languages. J. Evol. Biol. *26*, 1244–1253.
23. Janhunen, J. (2009). Proto-Uralic—what, where, and when? Suomalais-Ugrilaisen Seuran Toimituksia *258*, 57–78.
24. Kallio, P. (2006). On the Earliest Slavic Loanwords in Finnic. Slavica Helsingiensia *27*, 154–166.
25. Lang, V. (2016). Early Finnic-Baltic contacts as evidenced by archaeological and linguistic data. ESUKA-JEFUL *7*, 11–38.
26. Bjørnflaten, J.I. (2006). Chronologies of the Slavicization of Northern Russia Mirrored by Slavic Loanwords in Finnic and Baltic. In The Slavicization of the Russian North Mechanisms and Chronology, J. Nuorluoto, ed. (Helsinki: Department of Slavonic and Baltic Languages and Literatures, University of Helsinki), pp. 50–77.
27. Maurits, L., de Heer, M., Honkola, T., Dunn, M., and Vesakoski, O. (2020). Best practices in justifying calibrations for dating language families. J. Lang. Evol. *5*, 17–38.
28. Nevanlinna, H.R. (1972). The Finnish population structure. A genetic and genealogical study. Hereditas *71*, 195–236.
29. Peltonen, L., Jalanko, A., and Varilo, T. (1999). Molecular genetics of the Finnish disease heritage. Hum. Mol. Genet. *8*, 1913–1923.
30. Norio, R. (2003). Finnish Disease Heritage I: characteristics, causes, background. Hum. Genet. *112*, 441–456.
31. Saag, L., Varul, L., Scheib, C.L., Stenderup, J., Allentoft, M.E., Saag, L., Pagani, L., Reidla, M., Tambets, K., Metspalu, E., et al. (2017). Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe. Curr. Biol. *27*, 2185–2193.e6.
32. Lamnidis, T.C., Majander, K., Jeong, C., Salmela, E., Wessman, A., Moiseyev, V., Khartanovich, V., Balanovsky, O., Ongyerth, M., Weihmann, A., et al. (2018). Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. Nat. Commun. *9*, 5018.
33. Skoglund, P., Malmström, H., Omrak, A., Raghavan, M., Valdiosera, C., Günther, T., Hall, P., Tambets, K., Parik, J., Sjögren, K.G., et al. (2014). Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. Science *344*, 747–750.
34. Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M.T., Götherström, A., and Jakobsson, M. (2012). Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science *336*, 466–469.
35. Mittnik, A., Wang, C.C., Pfrengle, S., Daubaras, M., Zarina, G., Hallgren, F., Allmae, R., Khartanovich, V., Moiseyev, V., Torv, M., et al. (2018). The genetic prehistory of the Baltic Sea region. Nat. Commun. *9*, 442.
36. Martin, A.R., Karczewski, K.J., Kerminen, S., Kurki, M.I., Sarin, A.P., Artomov, M., Eriksson, J.G., Esko, T., Genovese, G., Havulinna, A.S., et al. (2018). Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. Am. J. Hum. Genet. *102*, 760–775.
37. Kerminen, S., Havulinna, A.S., Hellenthal, G., Martin, A.R., Sarin, A.P., Perola, M., Palotie, A., Salomaa, V., Daly, M.J., Ripatti, S., et al. (2017). Fine-Scale Genetic Structure in Finland. G3 *7*, 3459–3468.
38. Pankratov, V., Montinaro, F., Kushniarevich, A., Hudjashov, G., Jay, F., Saag, L., Flores, R., Marnetto, D., Seppel, M., Kals, M., et al. (2020). Differences in local population history at the finest level: the case of the Estonian population. Eur. J. Hum. Genet. *28*, 1580–1591.
39. Freyman, W.A., McManus, K.F., Shringarpure, S.S., Jewett, E.M., Bryc, K., Auton, A.; and 23 and Me Research Team (2021). Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows-Wheeler Transform. Mol. Biol. Evol. *38*, 2131–2151.
40. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

41. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
42. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *2010*, t5448.
43. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal *17*, 10–12.
44. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
45. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.; and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
46. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
47. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.
48. Fu, Q., Mittnik, A., Johnson, P.L.F., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., et al. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. Curr. Biol. *23*, 553–559.
49. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*, 356.
50. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. https://doi.org/10.1101/201178.
51. Mitt, M., Kals, M., Pärn, K., Gabriel, S.B., Lander, E.S., Palotie, A., Ripatti, S., Morris, A.P., Metspalu, A., Esko, T., et al. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur. J. Hum. Genet. *25*, 869–876.
52. Abraham, G., Qiu, Y., and Inouye, M. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics *33*, 2776–2778.
53. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
54. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. (Springer).
55. Karmin, M., Saag, L., Vicente, M., Wilson Sayres, M.A., Järve, M., Talas, U.G., Rootsi, S., Ilumäe, A.M., Mägi, R., Mitt, M., et al. (2015). A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. *25*, 459–466.
56. Poznik, G.D., Xue, Y., Mendez, F.L., Willems, T.F., Massaia, A., Wilson Sayres, M.A., Ayub, Q., McCarthy, S.A., Narechania, A., Kashin, S., et al.; 1000 Genomes Project Consortium (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat. Genet. *48*, 593–599.
57. Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinformatics *47*, 1–34.
58. Speed, D., and Balding, D.J. (2015). Relatedness in the post-genomic era: is it still useful? Nat. Rev. Genet. *16*, 33–44.
59. Free Software Foundation (2014). GNU Datamash. https://www.gnu.org/software/datamash/.
60. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput. Biol. *12*, e1004842.
61. Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal. Complex Syst. *1695*, 1–9.
62. Hothorn, T., and Hornik, K. (2019). exactRankTests: Exact Distributions for Rank and Permutation Tests. R package version 08-31 (The R Project).
63. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
64. Saag, L., Vasilyev, S.V., Varul, L., Kosorukova, N.V., Gerasimov, D.V., Oshibkina, S.V., Griffith, S.J., Solnik, A., Saag, L., D'Atanasio, E., et al. (2021). Genetic ancestry changes in Stone to Bronze Age transition in the East European plain. Sci. Adv. *7*, 1–17.
65. Chaitanya, L., Breslin, K., Zuñiga, S., Wirken, L., Pośpiech, E., Kukla-Bartoszek, M., Sijen, T., Knijff, P., Liu, F., Branicki, W., et al. (2018). The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Sci. Int. Genet. *35*, 123–135.
66. Ringbauer, H., Coop, G., and Barton, N.H. (2017). Inferring Recent Demography from Isolation by Distance of Long Shared Sequence Blocks. Genetics *205*, 1335–1351.
67. Browning, S.R., and Browning, B.L. (2015). Accurate Nonparametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum. Genet. *97*, 404–418.
68. Burger, J., Link, V., Blöcher, J., Schulz, A., Sell, C., Pochon, Z., Diekmann, Y., Žegarac, A., Hofmanová, Z., Winkelbach, L., et al. (2020). Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong Selection over the Last 3,000 Years. Curr. Biol. *30*, 4307–4315.e13.
69. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
70. Lamason, R.L., Mohideen, M.A.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science *310*, 1782–1786.
71. Basu Mallick, C., Iliescu, F.M., Möls, M., Hill, S., Tamang, R., Chaubey, G., Goto, R., Ho, S.Y.W., Gallego Romero, I., Crivellaro, F., et al. (2013). The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. PLoS Genet. *9*, e1003912.
72. Norton, H.L., Kittles, R.A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V.A., Bradley, D.G., McEvoy, B., and Shriver, M.D. (2007). Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol. Biol. Evol. *24*, 710–722.

73. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503.
74. Davies, R.W., Flint, J., Myers, S., and Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. Nat. Genet. *48*, 965–969.
75. Monroy Kuhn, J.M., Jakobsson, M., and Günther, T. (2018). Estimating genetic kin relationships in prehistoric populations. PLoS ONE *13*, e0195491.
76. Martin, M.D., Jay, F., Castellano, S., and Slatkin, M. (2017). Determination of genetic relatedness from low-coverage human genome sequences using pedigree simulations. Mol. Ecol. *26*, 4145–4157.

The American Journal of Human Genetics, Volume *108*

Supplemental information

Patterns of genetic connectedness between modern and medieval Estonian genomes reveal the origins of a major ancestry component of the Finnish population

Toomas Kivisild, Lehti Saag, Ruoyun Hui, Simone Andrea Biagini, Vasili Pankratov, Eugenia D'Atanasio, Luca Pagani, Lauri Saag, Siiri Rootsi, Reedik Mägi, Ene Metspalu, Heiki Valk, Martin Malve, Kadri Irdt, Tuuli Reisberg, Anu Solnik, Christiana L. Scheib, Daniel N. Seidman, Amy L. Williams, Estonian Biobank Research Team, Kristiina Tambets, and Mait Metspalu

Supplemental Materials

Supplemental Figures

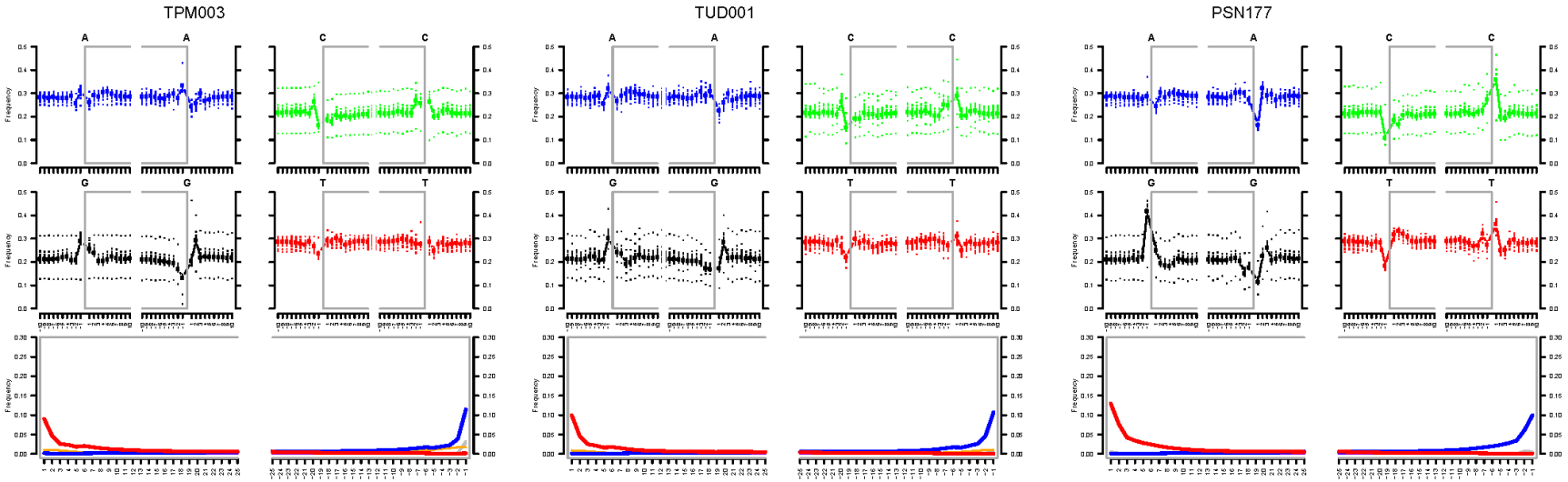


Figure S1. Genome-wide average damage rates in the three newly reported ancient genomes. The top plots show the frequency of each nucleotide in the 10 positions before and after each end of the DNA fragments that mapped to the human genome. The plots on the bottom show nucleotide substitutions at the 25 positions at each end of the DNA fragments that mapped to the human genome: the red line shows cytosine to thymine substitutions and the blue line guanine to adenine substitutions. Further details about the three genomes are reported in Table S1.

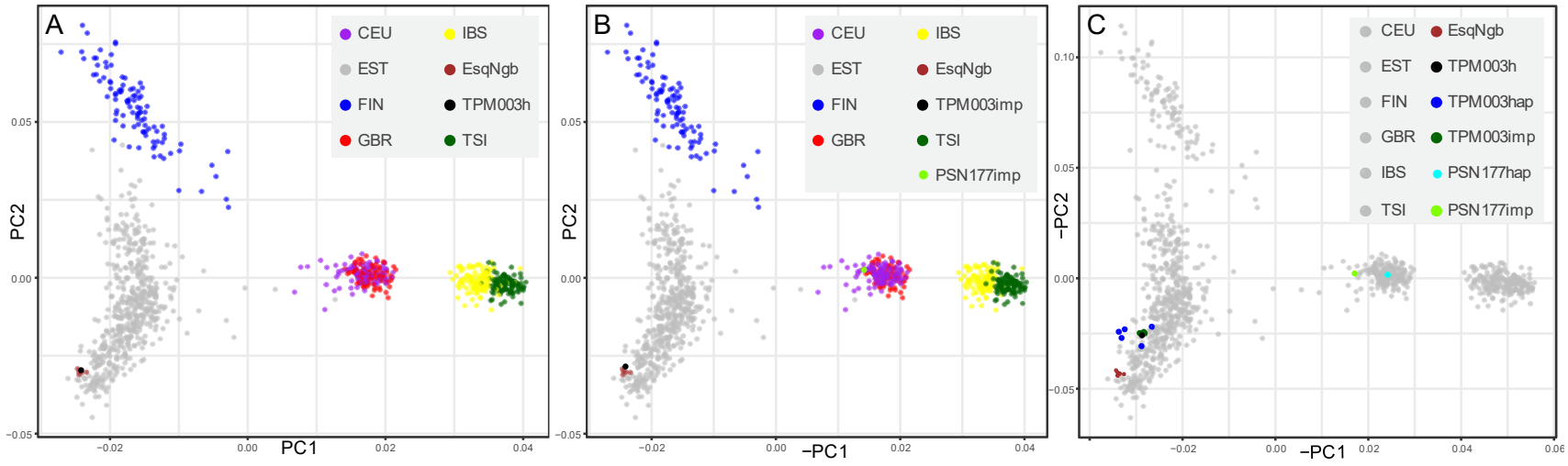


Figure S2. Smartpca analyses of TPM003 in context of 526 EstBB Estonians and 503 Europeans from 1000 GP. Genotypes were first directly called from the 23x coverage sample (TPM003h) without imputation. **A.** The placement of TPM003h on the PC plot is shown without projection. **B.** The placement on the PC plot of the down-sampled (to 0.1x) replica TPM003imp which was imputed and included to smartpca analyses without projection. **C.** PCA plot in which TPM003h is shown projected on PC-s calculated only from the modern samples along with projections of the haploid called (blue, TPM003hap) and imputed (green, TPM003imp) versions of its five independently down-sampled (0.1x) replicas. Through each plot the EstBB closest Euclidean squared distance neighbors (EsqNgb) of TPM003h by PC1 and PC2 coordinates are shown in brown.

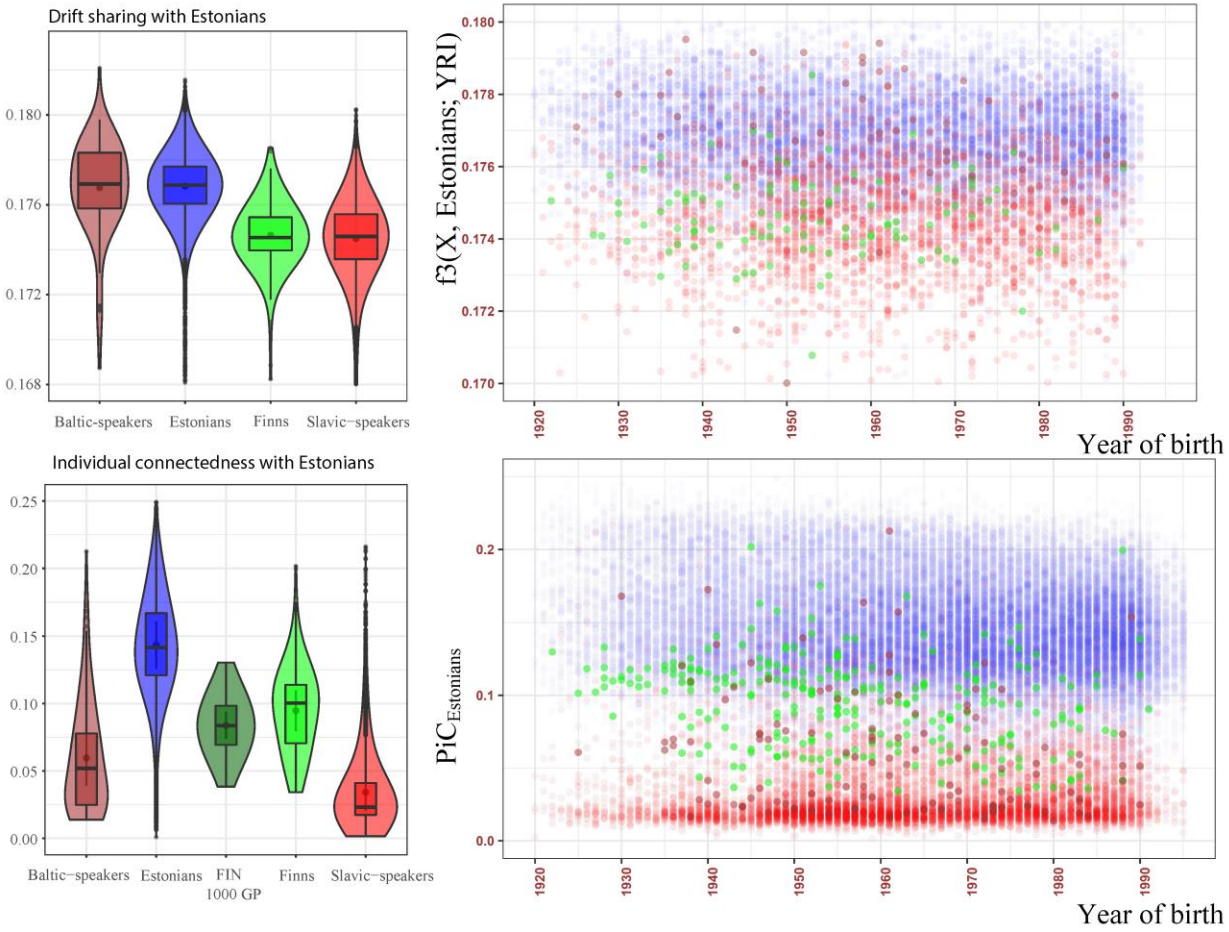


Figure S3. Drift sharing and individual connectedness with present-day Estonians. Upper panel shows drift sharing assessed using outgroup f3 statistics on Estonian, Finnish, Baltic and Slavic speaking subset of 47,015 EstBB individuals whose birthplace and year of birth data was available. Lower panel shows individual connectedness (PiC) with Estonians in the same subset of EstBB individuals as in the upper panel. The y axis shows the proportion of Estonians with whom the given individual shares at least one LSAI segment longer than 5cM and kinship coefficient >0.0005 (equivalent to the average expectations for 10th degree of genetic relatedness in large populations).

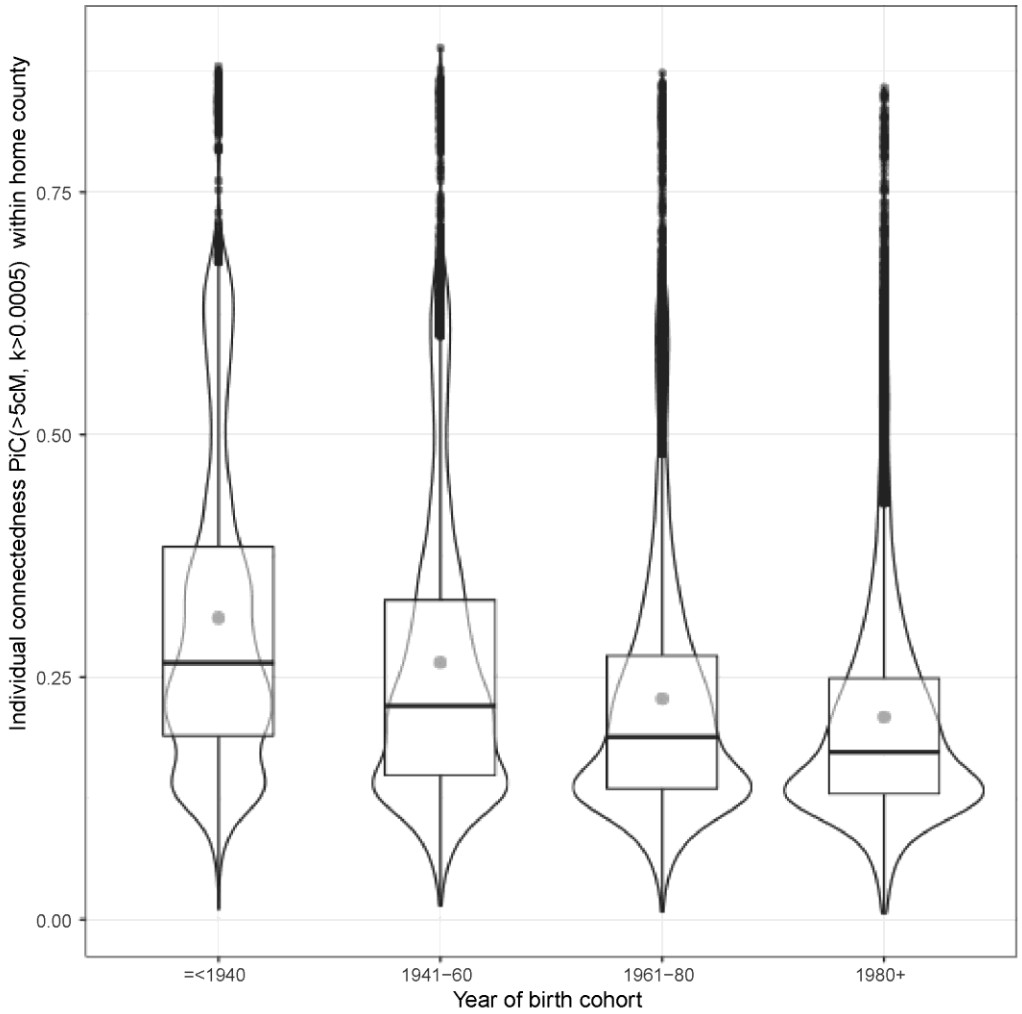


Figure S4. Proportions of individual connectedness of 47,015 EstBB samples (with birth place and year of birth information) within their home county by year of birth cohorts: (1) =<1940 cohort with 4,370 individuals, (2) 1941-60 cohort with 11,869 individuals, (3) 1961-80 cohort with 18,101 individuals, and (4) 1980+ cohort with 13,654 individuals.

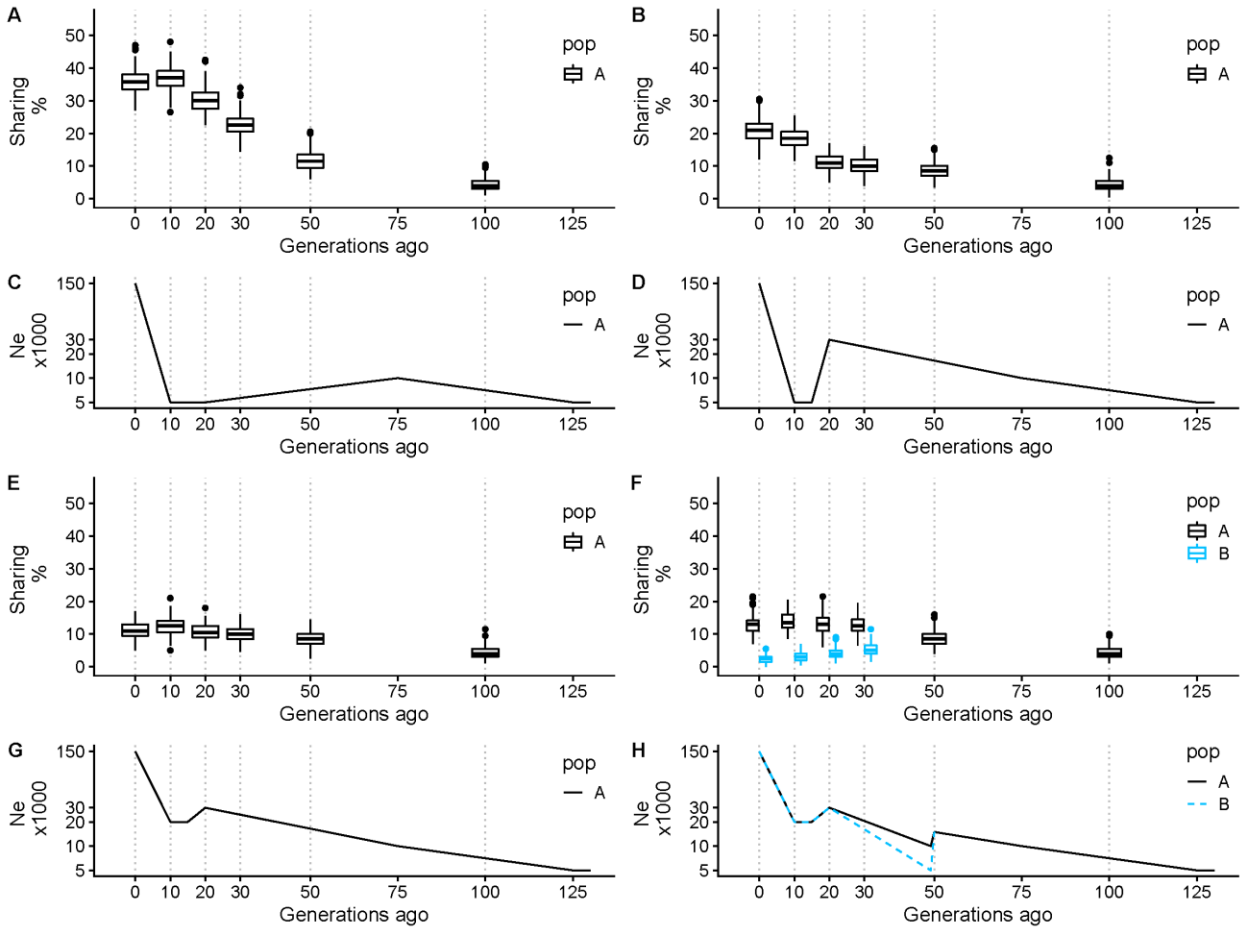


Figure S5. Individual connectedness among modern and ancient samples simulated under demographic scenarios inferred from whole genome sequences of present day Estonians (Pakratov et al. 2020). Panels A, B, E and F show the percentage of samples (PiC score) from different time points sharing at least one LSAI segment of 5 cM or longer and having a kinship coefficient above 0.0005 with each contemporary individual from population A. Panels C, D, G and H show the corresponding demography simulated. In panel H population B split from population A 50 generations ago. Y-axis in C, D, G and H is logarithmic. Grey vertical lines show the time points of sampling. Demographic scenarios C, D and G correspond to simplified representations of the demographic history of South-East, South-West and North-West Estonia [1]. Note the simulations are based on genome size approximately corresponding to 1/5th of the size of the human genome and therefore the presented sharing % is not exactly comparable to estimates based on the empirical results.

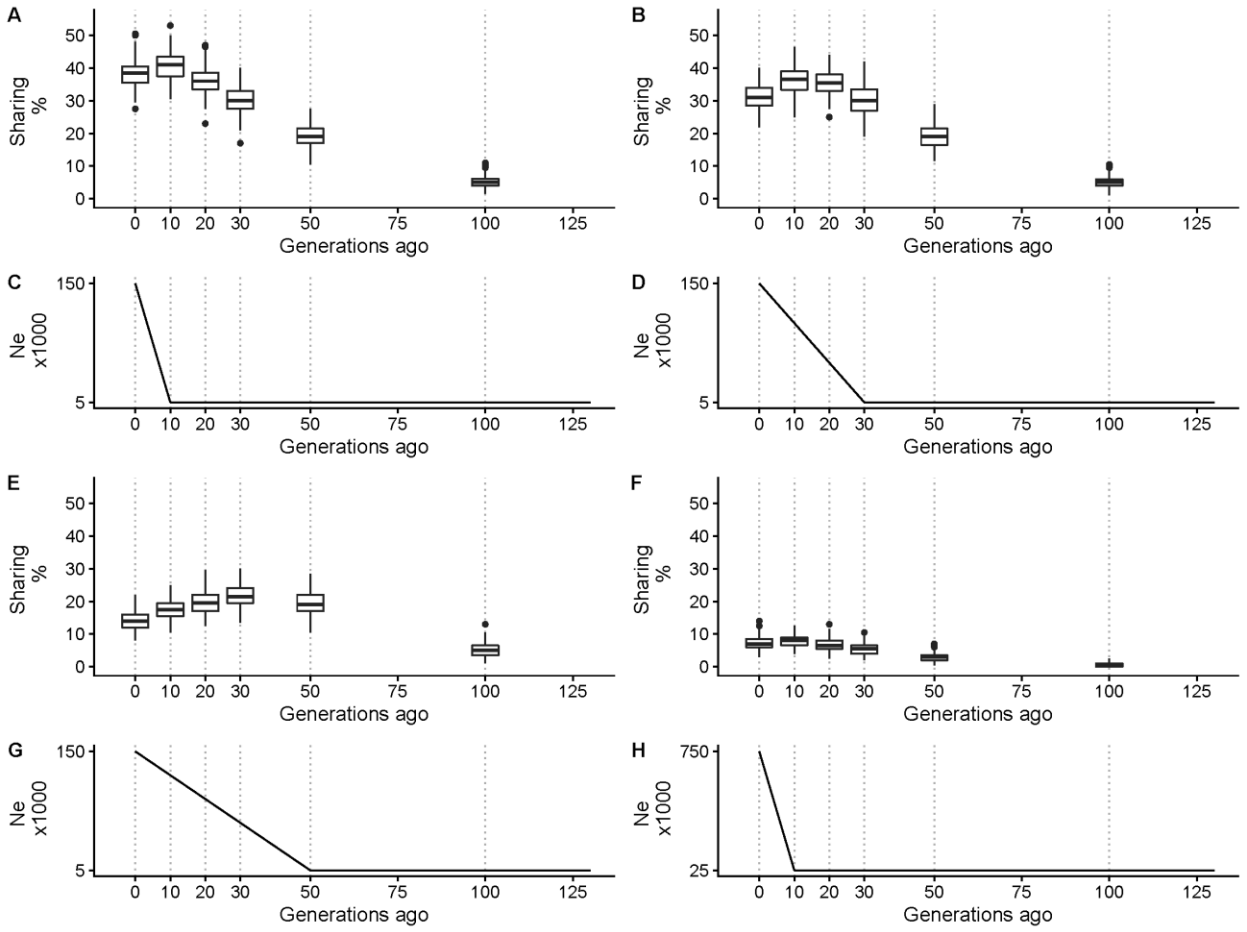
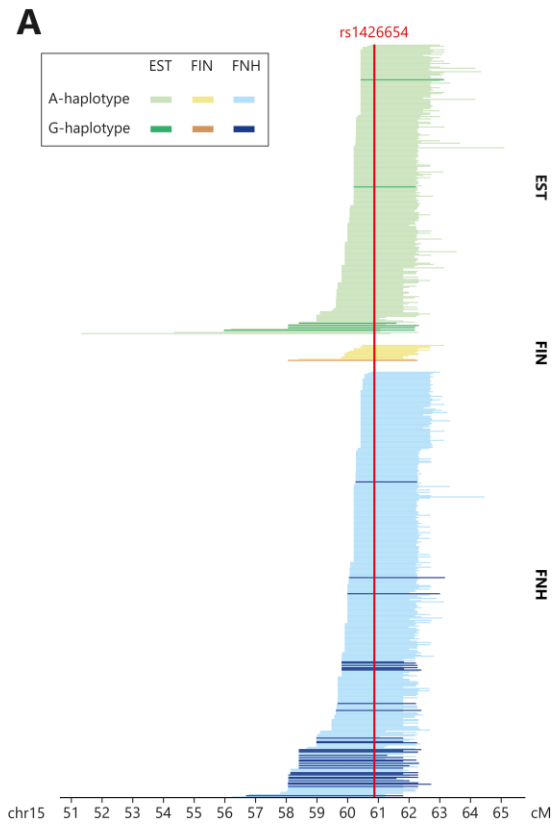


Figure S6. Individual connectedness among modern and ancient samples simulated under various demographic growth scenarios. Panels A, B, E and F show the percentage of samples (PiC score) from different time points sharing at least one IBD segment of 5 cM or longer and having a kinship coefficient above 0.0005 with each contemporary individual. Panels C, D, G and H show the corresponding demography simulated. In panel H population B split from population A 50 generations ago. Y axis in C, D, G and H is logarithmic. Grey vertical lines show the time points of sampling.



B

Region	Population	N	rs1426654-G	TPM003-shared	Reference
Europe	EST	2076	0.005	0.364	EST
	FNH	1918	0.022	0.452	HRC
	CEU	99	0.000	-	1000 GP
	GBR	91	0.000	-	1000 GP
	IBS	107	0.000	-	1000 GP
	FIN	99	0.010	0.500	1000 GP
	TSI	107	0.005	0.000	1000 GP
East Asia	CDX	93	1.000	0.000	1000 GP
	CHB	103	0.971	0.000	1000 GP
	CHS	105	0.981	0.000	1000 GP
	JPT	104	0.995	0.010	1000 GP
	KHV	99	0.995	0.000	1000 GP
South Asia	BEB	86	0.465	0.000	1000 GP
	GIH	103	0.049	0.000	1000 GP
	ITU	102	0.348	0.000	1000 GP
	PJL	96	0.219	0.000	1000 GP
	STU	102	0.515	0.000	1000 GP
Africa	ACB	96	0.896	0.000	1000 GP
	ASW	61	0.811	0.000	1000 GP
	ESN	99	0.975	0.000	1000 GP
	GWD	113	0.925	0.000	1000 GP
	LWK	99	0.924	0.000	1000 GP
	MSL	85	0.912	0.000	1000 GP
	YRI	108	0.986	0.000	1000 GP
America	CLM	94	0.282	0.000	1000 GP
	MXL	64	0.484	0.000	1000 GP
	PEL	85	0.718	0.000	1000 GP
	PUR	104	0.231	0.000	1000 GP

Figure S7: LSAI sharing between medieval TPM003 and modern Estonians and Finns in the *SLC24A5*. **A.** LSAI segments (> 2cM) shared between TPM003 and Estonians (EST, 2076 individuals from the high-coverage Estonian reference panel), Finns from 1000 GP (FIN) and Finns from the HRC (FNH). Segments in samples with the G allele at rs1426654 are in darker shades. Other tested Europeans populations from the 1000 GP that did not share G-segments with TPM003 are not shown. **B.** Worldwide frequency distribution of the G allele of rs1426654 (rs1426654-G) and relative frequency of the 200kb core G-haplotype specifically observed in TPM003 (TPM003-shared).

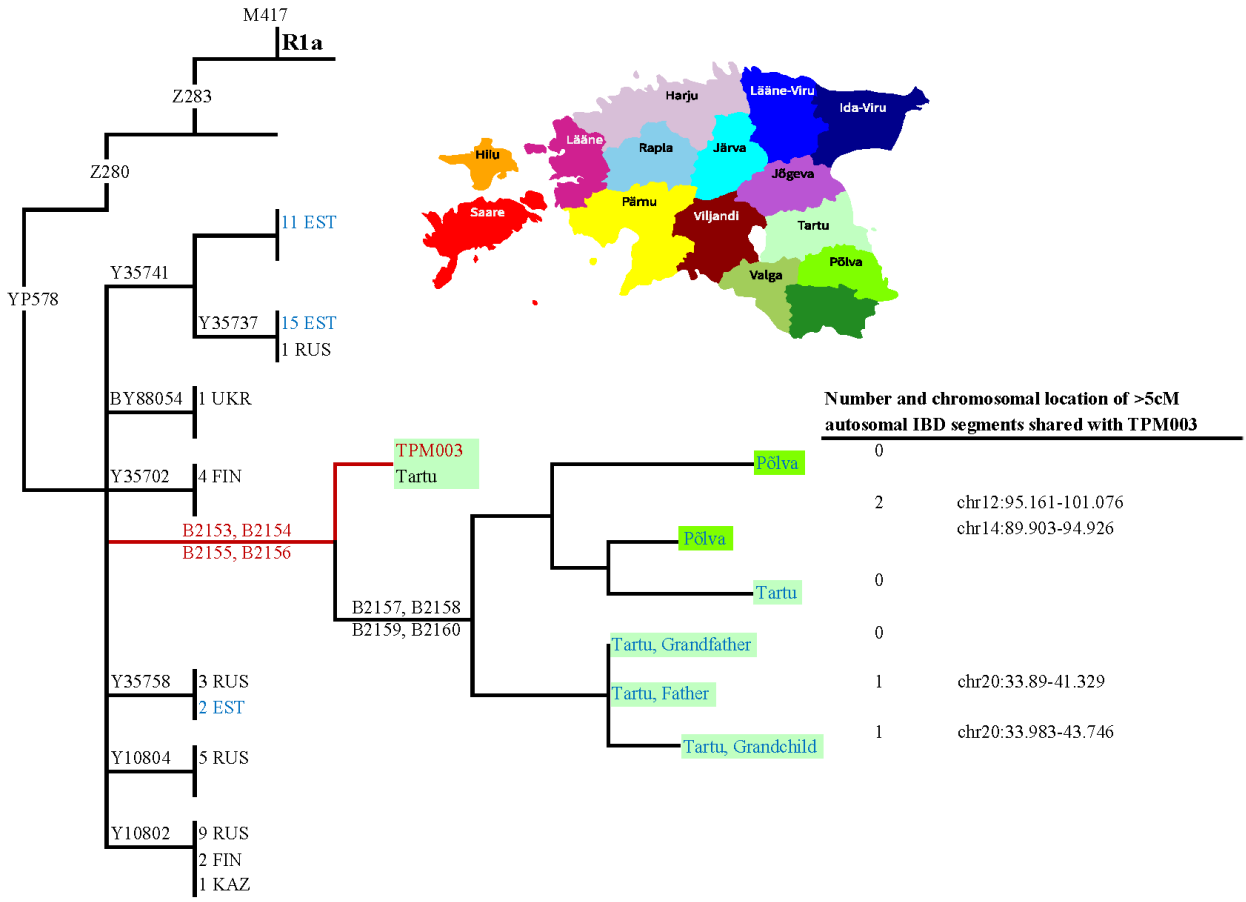


Figure S8. The placement of the medieval sample TPM003 in the context of phylogenetic tree of Y chromosome haplogroup R1a-YP578 in modern samples and his autosomal IBD sharing with present-day Estonians. Y chromosome lineages reported in the Y-Full YTree are shown in black and individuals from the Estonian Biobank in blue font. The birthplaces of the six Estonian Biobank samples phylogenetically closest to TPM003 are shown with colors corresponding to colors in the county map of Estonia.

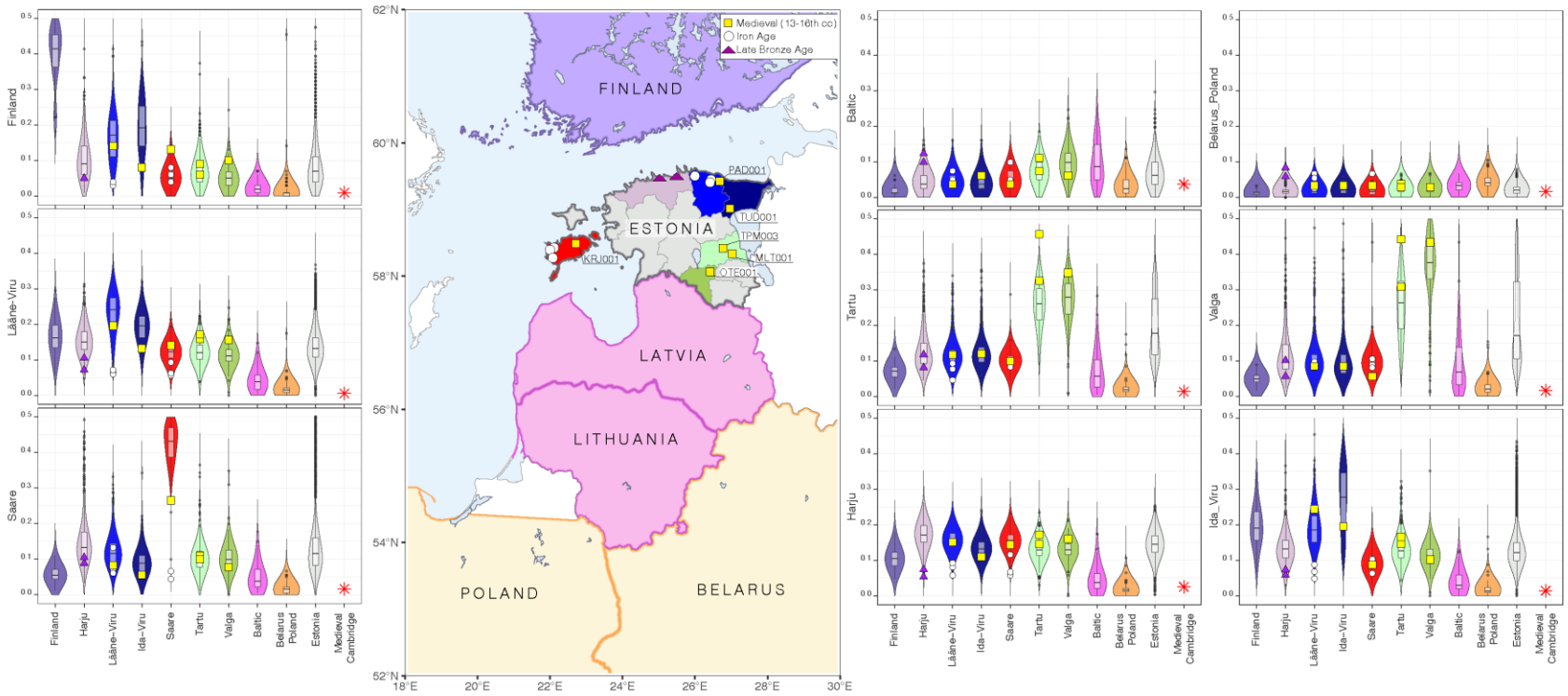


Figure S9. Genetic connectedness with nine target populations in modern populations of the circum-Baltic region and Estonian Bronze, Iron Age and medieval genomes. Each individual violin plot shows the distribution of PiC scores which reflect individual probabilities of >5 cM LSAI sharing and kinship coefficient estimate higher than 0.0005 with individuals from the population shown on the y axis. Distributions of the present-day populations are shown by the colors according to the map. Present-day genomes include 99 Finnish individuals (FIN) from the 1000 Genomes Project data, 1,880 Estonian Biobank Estonians born before 1940 in Harju, Saare, Viru, Tartu, and Valga counties, and 320 Estonian Biobank Latvians, Lithuanians, Belarus and Polish individuals born outside Estonia. All ancient samples, shown with squares (medieval), circles (Iron Age), and triangles (Bronze Age) have been imputed, including one medieval 0.1x coverage British genome - as a control, using 2,092 Estonian high coverage sequences as a reference panel.

Supplemental Tables

Heterozygote call accuracy

MAF	DS1-0.1x	DS2-0.1x	DS3-0.1x	DS4-0.1x	DS5-0.1x	DS6-0.3x
<0.01	0.7537	0.7448	0.7376	0.7467	0.7471	0.8197
0.01-0.05	0.9468	0.9446	0.9424	0.9421	0.9449	0.9644
0.05-0.1	0.9753	0.9747	0.9715	0.9756	0.9756	0.9863
0.1-0.3	0.9856	0.9858	0.9843	0.9835	0.9857	0.9929
>0.3	0.9896	0.9896	0.9901	0.9895	0.989	0.9947
CV	0.9866	0.9867	0.986	0.9857	0.9864	0.9932

Proportion of heterozygote positions retained

MAF	DS1-0.1x	DS2-0.1x	DS3-0.1x	DS4-0.1x	DS5-0.1x	DS6-0.3x
<0.01	0.5096	0.4981	0.4882	0.4993	0.5009	0.6776
0.01-0.05	0.7402	0.7364	0.7336	0.7343	0.7377	0.8777
0.05-0.1	0.8469	0.8441	0.8364	0.8443	0.8472	0.9405
0.1-0.3	0.8696	0.8671	0.8649	0.8660	0.8684	0.9603
>0.3	0.8733	0.8698	0.8680	0.8685	0.8661	0.9626
CV	0.8693	0.8664	0.8639	0.8653	0.8654	0.9596

Table S2. Heterozygote genotype imputation accuracy estimates by five minor allele frequency (MAF) classes for a medieval 23x coverage genome TPM003 and its five down-sampled replicas DS1-5 at 0.1x and one, DS6, at 0.3x coverage.

Note: Accuracy estimates reflect the proportion of heterozygous sites of the high coverage genome that were imputed as heteroz. CV - common variants, MAF >0.05.

ID		Period	Age	Geographic origin	EST	FIN	BAL	EUR
Average		**Present-day**		**FIN**	**0.087**	**0.523**	**0.020**	**0.003**
Average		**Present-day**		**Latvia and Lithuania**	**0.059**	**0.020**	**0.098**	**0.003**
Average		**Present-day**		**Estonia**	**0.137**	**0.087**	**0.059**	**0.003**
Average		**Medieval**		**Estonia**	**0.163**	**0.101**	**0.064**	**0.003**
MAL001		Medieval	435	Mäletjärve, Tartumaa	36738	6	9	0
TUD001		Medieval	520	Tudulinna, Ida-Virumaa	13709	8	5	4
TPM003		Medieval	570	Tartu, Tartumaa	28188	9	6	0
OTE001		Medieval	640	Otepää, Valgamaa	31719	10	5	1
KRJ001		Medieval	705	Karja, Saaremaa	15880	13	3	1
PAD001		Medieval	815	Pada, Lääne-Virumaa	14538	14	3	1
Average		**Iron Age**		**Estonia**	**0.068**	**0.047**	**0.058**	**0.004**
KUR001	V12	Iron Age	2220	Kurevere, Saaremaa	10812	4	5	0
KUR002	V11	Iron Age	2315	Kurevere, Saaremaa	8655	6	4	1
LOO001	X04	Iron Age	2440	Loona, Saaremaa	10092	8	7	2
VOT001	VII4	Iron Age	2600	Võhma, Lääne-Virumaa	10351	4	3	2
KND001	V10	Iron Age	2630	Kunda, Lääne-Virumaa	9021	3	4	0
KND002	0LS10	Iron Age	2685	Kunda, Lääne-Virumaa	10044	3	5	4
Average		**Bronze Age**		**Estonia**	**0.064**	**0.051**	**0.093**	**0.005**
REB002	X14	Bronze Age	2625	Rebala, Harju	9661	5	7	2
MUL001	X08	Bronze Age	2890	Muuksi, Harju	8696	5	8	2

Table S3. Time series of IBD sharing of Estonian Bronze, Iron Age and medieval genomes with 143,774 modern Estonians and 1000 Genome Project Europeans.

Note for each ancient sample the number of present-day genomes is reported with whom they share at least one >5cM IBD segment and kinship coefficient >0.0005. IBD analyses were performed with IBIS on 254326 SNPs with MAF>0.05 using the -setIndexEnd option for each ancient individual separately using only variants for which that individual had genotype imputed with GP>0.99. EST - 143,774 Estonian biobank individuals, FIN - 99 Finns, EUR - 404 individuals including 99 CEU, 91 GBR, 107 TSI and 107 IBS from the 1000 Genomes project data. BAL - 81 Baltic speakers (Latvians, Lithuanians) of EstBB.

	N	1	81	99	2	6	6	4419
	threshold	MedUK	Baltic	FIN	BrAgeEST	IrAgeEST	MedEST	EstBB
EstBB	2cM	0.242	0.396	0.444	0.672	0.623	0.615	0.553
	5cM	0.015	0.069	0.085	0.082	0.080	0.201	0.185
	7cM	0.005	0.040	0.046	0.023	0.032	0.127	0.130
	10cM	0.000	0.009	0.007	0.000	0.015	0.020	0.047
FIN	2cM	0.182	0.219	0.911	0.409	0.507	0.452	0.444
	5cM	0.010	0.020	0.526	0.051	0.047	0.097	0.085
	7cM	0.010	0.009	0.326	0.010	0.019	0.055	0.046
	10cM	0.000	0.002	0.093	0.000	0.003	0.006	0.007
Baltic	2cM	0.235	0.462	0.219	0.735	0.572	0.485	0.396
	5cM	0.037	0.098	0.020	0.082	0.070	0.065	0.069
	7cM	0.012	0.058	0.009	0.023	0.031	0.039	0.040
	10cM	0.000	0.015	0.001	0.000	0.000	0.009	0.009

Table S4. Average proportions of IBD sharing among population for four different IBD segment length thresholds.

Note - IBD inferences were made with IBIS using kinship coefficient 0.0005 threshold in addition to the IBD segmenth length threshold.

BrAgeEST, IrAgeEST and MedEST refer to pools of Bronze Age, Iron Age and Medieval genomes from Estonia, respectively.

MedUK - medieval (14th-15th cc) 0.1x coverage genome from StJohn's Hospital, Cambridge, UK

EstBB samples involve only 4419 individuals born before 1940 in Estonia; FIN - 99 Finnish samples from the 1000 Genomes Project,

Baltic - Baltic speakers (Latvians and Lithuanians) of the EstBB.

Parameters	Segment length: SNPs	7-10 cM PPV	7-10 cM TPR	15-20 cM PPV	15-20 cM TPR	≥20 cM PPV	≥20 cM TPR
MAF ≥ 0.005, -maxDist 0.12, default -mt	611,324	0.931	0.947	0.969	0.991	0.991	0.996
MAF ≥ 0.02, -maxDist 0.12, default -mt	449,526	0.939	0.927	0.972	0.989	0.992	0.995
MAF ≥ 0.05, -maxDist 0.12, -mt 300	312,685	0.956	0.878	0.979	0.982	0.994	0.993
MAF ≥ 0.05, -maxDist 0.2, -mt 300	312,685	0.956	0.899	0.979	0.983	0.994	0.993

Table S5. The precision and sensitivity of IBIS to detect IBD/LSAI segments of different size from simulated data.

Note, the estimates are based on simulated data of related individuals using the UK Biobank data as input. SNPs with missingness ≥ 0.01 in the UKB genotype data were excluded. PPV - positive predictive value or precision, the fraction of the total length of all inferred IBD segments (within a given length bin) that overlap any true segment of any size. TPR - true positive rate or sensitivity, the fraction of the total length of all true IBD segments (in a length bin) that an algorithm calls as identical by descent, considering all called segments of any size.

ID1	depth	ID2	depth	IBD2	totalIBD	Segs	DR
Imputed low coverage vs directly called high coverage genome							
TPM003d6_i	0.3x	TPM003_h	23x	0.926	0.965	156	0
TPM003d1_p	0.1x	TPM003_h	23x	0.377	0.688	309	1
TPM003d1_i	0.1x	TPM003_h	23x	0.330	0.659	303	1
TPM003d3_i	0.1x	TPM003_h	23x	0.300	0.649	290	1
TPM003d2_i	0.1x	TPM003_h	23x	0.298	0.641	292	1
TPM003d5_i	0.1x	TPM003_h	23x	0.287	0.641	279	1
TPM003d4_i	0.1x	TPM003_h	23x	0.283	0.634	275	1
Two independent imputations from different downsampled replicas							
TPM003d1_p	0.1x	TPM003d6_i	0.3x	0.355	0.677	293	1
TPM003d1_i	0.1x	TPM003d6_i	0.3x	0.307	0.649	296	1
TPM003d2_i	0.1x	TPM003d6_i	0.3x	0.300	0.643	294	1
TPM003d3_i	0.1x	TPM003d6_i	0.3x	0.297	0.647	288	1
TPM003d5_i	0.1x	TPM003d6_i	0.3x	0.287	0.641	280	1
TPM003d4_i	0.1x	TPM003d6_i	0.3x	0.264	0.625	275	1
TPM003d1_p	0.1x	TPM003d5_i	0.1x	0.205	0.598	243	1
TPM003d1_p	0.1x	TPM003d2_i	0.1x	0.205	0.592	224	1
TPM003d1_p	0.1x	TPM003d3_i	0.1x	0.202	0.598	247	1
TPM003d1_i	0.1x	TPM003d2_i	0.1x	0.186	0.584	234	1
TPM003d4_i	0.1x	TPM003d5_i	0.1x	0.185	0.590	223	1
TPM003d2_i	0.1x	TPM003d5_i	0.1x	0.185	0.584	230	1
TPM003d1_i	0.1x	TPM003d3_i	0.1x	0.184	0.585	239	1
TPM003d1_p	0.1x	TPM003d4_i	0.1x	0.180	0.584	210	1
TPM003d1_i	0.1x	TPM003d5_i	0.1x	0.177	0.584	224	1
TPM003d1_i	0.1x	TPM003d4_i	0.1x	0.174	0.579	210	1
TPM003d2_i	0.1x	TPM003d3_i	0.1x	0.171	0.579	217	1
TPM003d3_i	0.1x	TPM003d5_i	0.1x	0.171	0.584	214	1
TPM003d3_i	0.1x	TPM003d4_i	0.1x	0.170	0.578	217	1
TPM003d2_i	0.1x	TPM003d4_i	0.1x	0.164	0.569	199	1
Two independent imputations from the same low coverage source							
KRJ001_p	0.9x	KRJ1001_i	0.9x	0.995	0.997	35	0
TPM003d1_p	0.1x	TPM003d1_i	0.1x	0.605	0.800	377	1
TUD001_p	0.1x	TUD001_i	0.1x	0.499	0.748	313	1

Table S6. Estimates of IBD recovery from imputed low coverage ancient genomes sampled from the same individual source.

Note, the imputations of the TPM003 sample were performed from 5 independently downsampled 0.1x replicas (d1-d5). samples imputed in a pool of 7 medieval low coverage genomes are shown with _p, those imputed independently with _i suffix. _h suffix refers to 23x TPM genome genotypes of which were called directly without imputation. totalIBD proportion is reported as twice the kinship coefficient

estimated with IBIS, IBD2 - proportion of genome shared in IBD2. IBD analyses were performed with IBIS with 5cM threshold on 217,554 SNPs with MAF >0.05 in Estonian Biobank data that had no missingness in any of the imputed and directly called in these ancient samples. 'Segs' - number of IBD segments detected

DR - degree of relatedness.

				match probability	
Inference	**h23x**	**i0.1x**	**overlap**	**m-all**	**m-overlap**
Number of EstBB samples with >5cM IBD sharing and kinship coefficient >0.0005 in both h23x and i0.1x	28987	28188	23739	0.933	0.842
Number of EstBB samples with >5cM IBD sharing and kinship coefficient >0.001 in both h23x and i0.1x	8663	8167	6333	0.971	0.775
Number of EstBB samples with >5cM IBD sharing and kinship coefficient >0.001 in i0.1x and with >5cM IBD sharing and kinship coefficient >0.0005 in high coverage TPM3	28987	8167	7849	0.998	0.961

Table S7. Match probabilities of LSAI sharing of TPM003 high coverage (h23x) and imputed 0.1x coverage (i0.1x) copies with 143,774 modern Estonian Biobank (EstBB) samples.

Note - match probability for all (m-all) is estimated as a proportion of matching inferences made for the entire pool of 143,774 individuals, including cases where both h23x and i0.1x have no relationship with modern samples (1-(i0.1x-overlap)/143774). Match probability for the overlap (m-overlap) represents the proportion of positive inferences of i0.1x IBD sharing with EstBB samples that match with h23x (overlap/i0.1x).

	Probabilities of individual connectedness with present-day populations						
	h23x	i0.1xp1	i0.1xs1	i0.1xs2	i0.1xs3	i0.1xs4	i0.1xs5
average depth	23x	0.1x	0.1x	0.1x	0.1x	0.1x	0.1x
Pop\imputation	none	pooled	separately	separately	separately	separately	separately
Poland/Belarus	0.029	0.038	0.075	0.046	0.046	0.05	0.05
Latvia/Lithuania	0.074	0.074	0.148	0.074	0.111	0.111	0.111
FIN (1000 GP)	0.071	0.091	0.121	0.091	0.081	0.111	0.101
Harju	0.136	0.145	0.204	0.155	0.169	0.196	0.153
Hiiu	0.069	0.111	0.111	0.153	0.153	0.097	0.111
Ida Viru	0.14	0.136	0.158	0.14	0.145	0.145	0.163
Jarva	0.171	0.193	0.182	0.193	0.193	0.171	0.193
Jogeva	0.168	0.168	0.203	0.19	0.211	0.18	0.16
Laane	0.145	0.145	0.205	0.145	0.193	0.157	0.145
Lääne Viru	0.145	0.16	0.169	0.181	0.187	0.166	0.151
Parnu	0.19	0.19	0.212	0.215	0.249	0.21	0.167
Polva	0.436	0.432	0.432	0.481	0.427	0.448	0.44
Rapla	0.167	0.167	0.132	0.149	0.211	0.175	0.158
Saare	0.11	0.11	0.126	0.126	0.121	0.104	0.099
Tartu	0.315	0.327	0.357	0.353	0.331	0.337	0.345
Valga	0.327	0.311	0.354	0.377	0.346	0.35	0.339
Viljandi	0.227	0.236	0.295	0.268	0.27	0.234	0.243
Voru	0.456	0.445	0.472	0.463	0.47	0.463	0.465
Correlation	1	0.995	0.974	0.983	0.984	0.992	0.989

Table S8. Probabilities of individual connectedness (PiC) of TPM003 and his imputed low coverage copies (i0.1x).

Note, h23x - TPM003 high coverage genome was used as a reference for calculating the correlations over presented PiC scores. IBD sharing between TPM003 and modern samples was estimated with IBIS using >5cM and kinship coefficient >0.0005 thresholds.

p1, s1-5 index numbers refer to five down-sampled copies of TPM003 generated with SAMTOOLS using different seeds.

TPM003p1 was imputed in a pool of 7 medieval genomes while TPM3s1-5 were imputed separately.

Correlation - the Pearson product-moment coefficient calculated between h23x and the imputed replicas.

Model	Pop	G	Ne	r	Demography
1	A	0	150	0.3401	
		10	5	0	
2	A	0	150	0.1134	
		30	5	0	
3	A	0	150	0.068	
		50	5	0	
4	A	0	750	0.3401	
		10	25	0	
5	A	0	150	0.3401	South-East Estonia
		10	5	0	
		20	5	-0.0126	
		75	10	0.0139	
		125	5	0	
6	A	0	150	0.3401	South-West Estonia
		10	5	0	
		15	5	-0.3584	
		20	30	0.02	
		75	10	0.0139	
		125	5	0	
7	A	0	150	0.2015	North-West Estonia
		10	20	0	
		15	20	-0.0811	
		20	30	0.02	
		75	10	0.0139	
		125	5	0	
8	A	0	150	0.2015	
		10	20	0	
		15	20	-0.0811	
		20	30	0.0379	
		49	10	-0.4055	
		50	15	0.0162	
		75	10	0.0139	
		125	5	0	
	B	0	150	0.2015	
		10	20	0	
		15	20	-0.0811	
		20	30	0.0618	
		49	5	-1.0986	
		50	15	Pop B merges with A	

Table S9. Description of demographic models used for simulations.

Note, G - number of generations back in time; Ne - effective population size, in thousands of individuals; r - growth rate; Demography - choice of demographic parameters of regional population histories according to Pankratov et al. 2020 [1].

B37position	ANC	DER	TPM003	Marker
17980713	G	A	A	B2153
19332625	A	G	G	B2154
19527255	T	C	C	B2155
21886504	T	C	C	B2156
7144971	G	A	G	B2157
14405906	C	T	C	B2158
16237271	A	G	A	B2159
21680037	A	C	A	B2160

Table S11. Newly defined Y chromosome markers within the R1a-YP578 clade.

The placement of the markers on the branches of the Y chromosome phylogeny is shown in Figure S8.

ID	23x	i0.3x	i0.1x
Eye colour	Blue	Blue	Blue
P Eye	0.92	0.94	0.94
Hair colour	Brown	Brown	Brown
P Hair	0.51	0.50	0.50
Hair Shade	Light	Light	Light
P Shade	0.91	0.92	0.92
Final Hair colour prediction	Brown/Dark brown	Brown/Dark brown	Brown/Dark brown
Skin colour	Intermediate	Intermediate	Intermediate
P skin	0.99	0.95	0.99

Table S13. Pigmentation phenotype predictions for TPM003 high-coverage and imputed copies.

Note, P Eye, P Hair, P Shade and P Skin refer to the probabilities of the most supported phenotype category according to the HIrisPlex-Stool, https://hirisplex.erasmusmc.nl/pdf/hirisplex.erasmusmc.nl.pdf.

Chr	SNP	Pos (hg19)	REF	ALT	TPM003 haplotype
15	rs2469592	48401875	A	G	A
15	rs8038571	48402368	A	G	A
15	rs938505	48405895	C	T	C
15	rs55728404	48411805	T	G	T
15	rs2675346	48411821	C	T	C
15	rs2433354	48414969	C	T	C
15	rs2459391	48415068	A	G	A
15	rs2433356	48416360	G	A	G
15	rs8041370	48425379	G	A	G
15	**rs1426654**	**48426484**	**A**	**G**	**G**
15	rs1878188	48430423	C	G	C
15	rs2469597	48438269	C	T	C
15	rs2459394	48444748	A	T	T
15	rs149639137	48476672	T	C	T
15	rs2413887	48485926	T	C	T
15	rs9920281	48514309	G	A	A

Table S14. Allelic states of TPM003 for *SLC24A5* variants defining its core rs1426654-G haplotype.

Supplemental Methods

Estimation of the accuracy of imputation and its effect on LSAI inference

Detailed summary of the sequence data of all the ancient samples used in this study is provided in Table S1.

Genotype calling and coverage down-sampling of the high coverage genome

Subsampling (-s) option in 'SAMTOOLS view' was used to generate five copies of the TPM003 genome at 0.1x coverage. The genotypes of each were then independently imputed with an approach described below for low coverage genomes.

Genotype calling and imputation of the low coverage genomes

Imputation analyses were performed using the pipeline described in detail elsewhere [2] and an ethnically matched panel of 2076 high coverage sequences from Estonia [3]. In short, genotype likelihood calls from low coverage genomes were estimated with ATLAS [4] followed by the first imputation step with BEAGLE 4.1 -gl [5], a GP>0.99 filtering of SNPs followed by the second BEAGLE 5.0 -gt [6] imputation step and a final (optional) step of GP filtering at different thresholds (no GP filter; 0.5; 0.9; 0.99). We note that substitution of BEAGLE 4.1 with GLIMPSE [7] in the first imputation step would yield slightly improved heterozygote calling accuracy at the cost of a significant reduction of high confidence SNPs [2]. Considering the restricted number (265,227) of SNPs with MAF>0.05 in the Illumina GSA array data available for the Estonian Biobank and the dependence of the downstream analyses on available SNP numbers the BEAGLE 4.1 – BEAGLE 5.0 pipeline was used throughout the analyses.

IBD/LSAI analyses

IBD inference of up to 6th degree relatives from high-density and quality genotype data can be achieved at high accuracy through long (>7cM) IBD segments [8]; however, the genotype density of modern biobank genotype data is often substantially reduced when merged with ancient imputed genomes because of the accumulative effect of missing data and because variants with low minor allele frequency (MAF<0.05) have low imputation accuracy [2]. To estimate the effect of reduced genotype density on LSAI inference with IBIS we simulated the effect of different minor allele frequency filters on the UK Biobank data and observed that while the accuracy of LSAI inference from ca 300K SNPs with MAF>0.05 can be achieved at >0.95, the sensitivity is lowered (0.899) with 600K SNPs with MAF>0.005 threshold (Table S5).

We used Ped-sim [9] to simulate 1,000 pairs of first through sixth degree relatives and 500 pairs each of seventh and eighth degree relatives (Table S5). For this, we used a sex-specific genetic map [10], crossover interference modelling [11; 12], and drew pedigree founder samples from 39,485 UK Biobank samples. These samples include unrelated individuals (defined for the purposes of this simulation as sixth degree or more distant relatives) within 62,508 UK Biobank samples previously selected to include a mix of fourth degree or more distant relatives and geographically distributed samples (see Seidman et al. [8]). These data were previously phased with Eagle 2.4 [8; 13]. We ran IBIS on the simulated genotypes and calculated sensitivity and the positive predictive value (PPV) by comparing the inferred segments with the true segments as produced by Ped-sim. These metrics are defined fractionally, with sensitivity being the proportion of the true segments' length inferred as an IBD segment/LSAI by IBIS, and PPV the fraction of the inferred segments that overlap a true segment.

To investigate the accuracy of LSAI inference from imputed low-coverage ancient genomes further, we compared the extent of IBIS reported IBD shared between high-coverage TPM003 and its down-sampled and imputed replicas. We found that while IBD2, corresponding to genomic regions with genotype identity between two genomes, can be recovered from individuals whose genomes have been imputed from 0.1x coverage (i0.1x) only at >30% and total IBD (i.e., both IBD1 and IBD2 segments) at >60% rate, the sensitivity of IBD detection is not sufficiently high for accurate estimation of close degrees of

relatedness: for the copies of the same source genome we inferred 1st rather than 0th degree of relatedness (Table S6). We retain 92.6% of IBD2, 96.5% of total IBD and accurate identification of degree of relatedness, however, in case of i0.3x and self-sharing IBD inference of genomes imputed independently from the same 0.9x coverage source at >0.99 accuracy (Table S6). We also observe that a 0.1x down-sampled replica of the TPM003 genome imputed in a pool together with six other medieval genomes showed higher accuracy than the other five independently imputed replicas (Table S6).

While the 0.1x imputed genomes are not suitable for accurate detection of close (<2nd degree) relatives, the accuracy of distant relationship detection between medieval and present-day Estonian genomes, using 5cM length and kinship coefficient 0.0005 thresholds appears to be sufficiently good for regional ancestry mapping. We find that on average 84% of the individual links detected above the threshold for the high-coverage TPM003 sample are replicated in its 0.1x down-sampled replicas and that 96% of the links of the imputed 0.1x genome at kinship coefficient threshold 0.001 match the inferences made from the high-coverage genome at kinship coefficient threshold 0.0005 (Table S7).

Phenotype prediction concordance rate and analysis of the SLC24A5 region

We were able to obtain genotype information for a total of 90 SNPs in the high-coverage TPM003, with a concordance rate with its down-sampled and imputed copies of 98.9 % with 0.3x (1 error and two NAs) and of 100% with 0.1x. Among the SNPs considered, we also analysed a subset of 35 markers included in the HIrisPlex-S set [14], obtaining the allelic information for a total of 31 markers, sufficient to have a prediction of blue eyes, brown/dark brown hair and intermediate skin. Consistent with the high concordance rate in the genotypes, we obtained the same prediction also for the 0.3x and 0.1x copy of TPM003it (Table S13).

To explore the LSAI sharing in the *SLC24A5* gene between TPM003 and modern Europeans, we ran IBIS with the same commands reported above with a >2cM threshold and the -setIndexEnd option to analyse

the samples with the G allele at rs1426654 against all the others. In this analysis, we included Estonians from the Estonian reference panel, Finns from Haplotype Reference Consortium (HRC) [15] and European samples from 1000GP. We then plotted the LSAI segments shared between TPM003 and other modern populations using the R package “ggplot2”, after excluding those groups not showing any segments with the G allele at rs1426654 shared with TPM003 (Figure S7 panel A). To identify the exact haplotype with the G allele at rs1426654 carried by TPM003, we created Haploview [16] input files (.ped and .info format) using PLINK, restricting the analysis to 200 kb around the variant (core block) and including TPM003, Estonians and Finns sharing LSAI segments with it. We then visualised the blocks using the "Solid spine of LD" option and listing all the haplotypes regardless of their frequency. We then selected 16 SNPs able to distinguish between different blocks with the G allele at rs1426654 (Table S14) and produced a fasta file with their allele information in 1) 2076 Estonians (a subset from the high-coverage Estonian reference panel); 2) Finns from HRC ; 3) 1000 GP. The worldwide frequency of the 16 NSP haplotype in the 200kb region surrounding the G-allele of TPM003 was estimated as the number of matching haploid sequences divided by the total number of haploid sequences considered (Figure S7 panel B).

Estimates of imputation accuracy on LSAI and diachronic connectedness inference

To gain further insights into regional and diachronic connectedness of medieval Estonians and to test how accurately it can be inferred from imputed data, we estimated the proportions of IBD sharing to modern populations for: 1) the high-coverage (23x) TPM003 genome, the genotypes of which were inferred directly without imputation; and, 2) the five down-sampled (0.1x) replicas, which were imputed. Consistent with the results of IBD recovery between imputed copies of the same genome (Table S6), we observe higher consistency in regional LSAI sharing patterns between the high coverage genome and its down-sampled replica that was imputed in a pool of other samples rather than imputed separately (Table S8).

References

1. Pankratov, V., Montinaro, F., Kushniarevich, A., Hudjashov, G., Jay, F., Saag, L., Flores, R., Marnetto, D., Seppel, M., Kals, M., et al. (2020). Differences in local population history at the finest level: the case of the Estonian population. Eur J Hum Genet 28, 1580-1591.
2. Hui, R.Y., D'Atanasio, E., Cassidy, L.M., Scheib, C.L., and Kivisild, T. (2020). Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. Sci Rep-Uk 10.
3. Mitt, M., Kals, M., Parn, K., Gabriel, S.B., Lander, E.S., Palotie, A., Ripatti, S., Morris, A.P., Metspalu, A., Esko, T., et al. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet 25, 869-876.
4. Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., and Wegmann, D. (2018). ATLAS: Analysis Tools for Low-depth and Ancient Samples. biRxiv.
5. Browning, B.L., and Browning, S.R. (2016). Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98, 116-126.
6. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet 103, 338-348.
7. Rubinacci, S., Ribeiro, D.M., Hofmeister, R., and Delaneau, O. (2020). Efficient phasing and imputation of low-coverage 1 sequencing data using large reference panels. bioRxiv.
8. Seidman, D.N., Shenoy, S.A., Kim, M., Babu, R., Woods, I.G., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., et al. (2020). Rapid, Phase-free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification. Am J Hum Genet 106, 453-466.
9. Caballero, M., Seidman, D.N., Qiao, Y., Sannerud, J., Dyer, T.D., Lehman, D.M., Curran, J.E., Duggirala, R., Blangero, J., Carmi, S., et al. (2019). Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. Plos Genet 15, e1007979.
10. Bherer, C., Campbell, C.L., and Auton, A. (2017). Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. Nat Commun 8, 14994.

11. Campbell, C.L., Furlotte, N.A., Eriksson, N., Hinds, D., and Auton, A. (2015). Escape from crossover interference increases with maternal age. Nat Commun 6, 6260.
12. Housworth, E.A., and Stahl, F.W. (2003). Crossover interference in humans. Am J Hum Genet 73, 188-197.
13. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Y, A.R., H, K.F., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 48, 1443-1448.
14. Chaitanya, L., Breslin, K., Zuniga, S., Wirker, L., Pospiech, E., Kukla-Bartoszek, M., Sijen, T., de Knijff, P., Liu, F., Branicki, W., et al. (2018). The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Sci Int-Gen 35, 123-135.
15. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics 48, 1279-1283.
16. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263-265.