Title: 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[2], Anu Solnik[2], Christiana L. Scheib[2,4,9], Daniel N. Seidman[10], Amy L. Williams[10], Estonian Biobank Research Team[2], Kristiina Tambets[2,11], Mait Metspalu[2,11]

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 - these authors contributed equally to this work

* Correspondence: Toomas Kivisild, toomas.kivisild@kuleuven.be

Abstract

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 (23x) and evaluated the performance of its down-sampled replicas. We find that ancient genomes imputed from >0.1x coverage can be reliably used in principal component analyses without projection. By searching for long shared allele intervals (LSAI; 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 cc 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 due to 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 11.3x 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 0.1x 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 [19-21]; 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 (Figure 1E-F). Linguistic studies have suggested that the differentiation of the Finnic from Finno-Volgaic languages dates back to 3,000-4,000 years ago [22-24]. 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 Age [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 is likely to post-date 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 cc, point to a possible prehistoric migration event from North Estonia to Finland after the second wave of Slavic expansion (8th–10th cc) 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 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 [31]. Both Estonia and Finland show internally high level of sub-structure [31-33] 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 century which could account for local patterns of Finnish IBD sharing in Northeast Estonia [33]. 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 ≥ 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 [34], we refer to them as long shared allele intervals (LSAI).

Material and Methods

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

Illumina GSA array data for 707,385 SNVs genotyped in 150,415 individuals of the Estonian Biobank was merged with the 1000 Genomes Project (1000 GP) Phase 3 [35] data. After applying --maf 0.05 --geno 0.005, --mind 0.01 filters in PLINK-1.9.0 [36] data for 143,774 Estonian Biobank individuals (representing >10% of the total population of Estonia), 503 individuals of European ancestry (CEU, GBR, FIN, IBS, TSI) from the 1000 GP and 254,325 overlapping SNVs with genetic map coordinates (build 37) were retained 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, a medieval tooth sample, PSN177, from the cemetery of the Hospital of St John, Cambridge, the UK, was analysed 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 of 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, apical tooth roots were broken off or cut off with a drill and used 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, EDTA and proteinase K was added and the samples were left to digest on a rotating mixer at 20 °C for 72 hours 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 MWCO PES, Sartorius) and purified in large volume columns (High Pure Viral Nucleic Acid Large Volume Kit, Roche) using the MinElute PCR Purification Kit (QIAGEN).

Sequencing libraries were built using NEBNext DNA Library Prep Master Mix Set for 454 (E6070, New England Biolabs) and Illumina-specific adaptors [37] following established protocols [19; 37]. The samples were purified between steps using 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 using HGS Diamond Taq DNA polymerase (Eurogentec). Three verification steps were implemented to make sure library preparation was successful and to measure the concentration of dsDNA/sequencing libraries – fluorometric quantitation (Qubit, Thermo Fisher Scientific), parallel capillary electrophoresis (Fragment Analyser, Agilent Technologies) and qPCR.

DNA was sequenced using the Illumina NextSeq 500 platform with the 75 bp single-end method. Firstly, samples were multiplexed to gain low coverage data. Later, an additional 4 full runs of data were generated 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 using cutadapt 1.11 [38]. Sequences shorter than 28 bp were removed to avoid random mapping of sequences from other species. The sequences were mapped to reference sequence GRCh37 (hs37d5) using Burrows-Wheeler Aligner (BWA 0.7.12) [39] 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 [40]. 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 [41] and reads with mapping quality under 10 were filtered out with samtools 1.3.

Due to post-mortem degradation, aDNA can be distinguished from modern DNA by shorter fragments and a high frequency of cytosine deamination at the ends of sequences. The program mapDamage2.0 [42] was used 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 [43]. This included calling an mtDNA consensus sequence based on reads with mapping quality at least 30 and positions with at least 5x 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 using the option trimBases to trim 7 bases from the ends of reads. For male individuals, X chromosome contamination was also estimated using the two methods in the script contamination.R incorporated in ANGSD [44].

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 [45] using Build37 refererence and --genotyping_mode GENOTYPE_GIVEN_ALLELES, --output_mode EMIT_ALL_SITES and --alleles variant.list options. In total, 12.6 million single nucleotide variants that had minor allele frequency higher than 0.1% in a subset of 2076 high coverage whole genome sequences [46] were used in variant.list. Called variants were filtered with GQ>30, 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 Supplementary Methods.

Principal component analyses (PCA)

FlashPCA2 [47] was used to perform principal component analysis on high coverage and imputed ancient genomes in the context of 69,218 Estonian Biobank samples and 503 Europeans (CEU, GBR, FIN, IBS, TSI) from the 1000 GP data [48]. After merging genotype data of 15 ancient (including 14 Estonian and 1 British sample) and 69,713 modern individuals we thinned the data by excluding variants in linkage disequilibrium with PLINK [36] --indep-pairwise 1000 50 0.5 option and excluded recommended[47] 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 (23x) and its down-sampled (to 0.1x) and imputed replicas in PCA performed with FlashPCA2 and smartpca in analyses together with a sub-sample of 1040 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 0.1x 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 using 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 was converted to EIGENSTRAT format using the program convertf from the EIGENSOFT 7.2.0 package [49]. The results were plotted in R with ggplot2 [50].

Y chromosome analyses

In total, 113,217 haplogroup informative Y chromosome variants from regions that uniquely map to Y chromosome [46; 51; 52] were called as haploid from the BAM file of the high coverage TPM003 with --doHaploCall function in ANGSD [44]. Derived and ancestral allele as well as haplogroup annotations for each of the called variants were added using BEDTools 2.19.0 [53] 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 [46] and the phylogenetic tree of Yfull YTree v8.10.0.

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] version 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, 10 cM thresholds) and -maxDist 0.1 and -mt 300 parameters. The total number of SNPs used varied between 244643-254326 MAF>0.05 variants.

Although IBIS has the highest IBD inference accuracy for >7cM segments [16], we use >5cM 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 [54], particular, 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 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 all low coverage (<0.3X) samples were analysed individually (in order to avoid cumulative loss of SNP numbers) against the EstBB and 1000 GP data 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 Supplementary 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, the count of connected individuals from group Z was estimated from sorted IBIS .coef output file using linux *join* function to add group codes to individual identifiers and using *crosstab* function of datamash [55] to generate the table of counts, each of which being 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 patters of IBD sharing between contemporary and ancient samples expected under different demographic scenarios we simulated 8 different demographic scenarios described in Supplementary Table S9 using msprime [56]. In all simulations we used the discrete time Wright-Fisher model (model="dtwf) to simulate generations 0 to 1000 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 chromosome one recombination map or in the junction between the two maps as well as positions with

derived allele's frequency in the simulated dataset less than 5% to match the filtering scheme applied on empirical data. The LSAI segments were detected using IBIS [16] with the same thresholds (at least one >5cM 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 Louvain algorithm [17], that is implemented in the R library *igraph* [57]. 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 [58].

In our pipeline, the igraph algorithm detects first all possible level-one communities, 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 test of statistical significance before moving to the following cycle. We let this process to 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, together with a per community list of included individuals. At this point, using a series of custom scripts, all the community levels are combined 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 using 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 4852 were removed from the intensity score matrix as outliers, including 281/320 Slavic or Baltic speaking EstBB participants and 49/4419 ethnic Estonians (Table S15). Community membership proportions were then plotted as pie charts (Figure 6) in R using ggplot2 package.

Phenotype prediction analyses

We used vcftools [59] (--snp option) to extract the genotype information at 104 phenotype informative markers already analysed [60], after excluding 9 SNPs absent in our Estonian reference panel, from the high-coverage TPM003 and for its downsampled (0.1x and 0.3x) and imputed copies. The genotypes have been then filtered to keep only variants with GP ≥ 0.99 and recorded as the number of effective alleles using PLINK (--recode A –recode-allele option). For the HIrisPlex-S set for the pigmentation prediction [61], the genotype data were uploaded to the HIrisPlex-S webtool after reformatting using *merge* function in R to combine information from all informative SNPs. The results of the webtool have been interpreted 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 Supplementary Methods section.

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 cc male individual (TPM003) from Tartu Püha Maarja (St. Mary) parish cemetery (Estonia) to an average depth of 23x (Table S1). We determined the genotype calls of the high coverage genome of TPM003 first directly and then compared the results against genotype calls from five down-sampled copies of 0.1x coverage, each imputed using a panel of 2,076 Estonian high coverage sequences [46]. 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 with a notable accuracy drop to <95% and <80% in variants with MAF <0.05 and <0.01, respectively (Table S2). In further analyses we used only variants with MAF >0.05.

We next analysed 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 Estonian Biobank (EstBB) and 503 Europeans from the 1000 GP [35] using FlashPCA2 [47] and smartpca [49] (Figure 2, S1). We observed a clear distinction of Estonians from other European populations including Finns (Figure 2A). By Fst statistic, Estonians are more differentiated from neighbouring Finns than 1000 GP Italians from Tuscany are, for example, from the Iberians (Figure 2B). We found that all 14 imputed Bronze, Iron Age and medieval samples from Estonia cluster together with present-day Estonians approximately within the same broad geographic regions of Estonia where 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 (Figure 2A,C). 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 0.1x coverage, mapped next to the closest neighbours from the EstBB in the PCA constructed using 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 (>5cM) 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 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 Fst 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 neighbours by PiC requires, however, further scrutiny with regards to the time depth of these connections. Analysis of ancient genomes can provide answers 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 (12th-16th cc) than EstBB and Iron (p=0.02, 2-tailed t-test) or Bronze Age (p=0.017) samples using LSAI length >5cM threshold. These results stand in contrast to the lack of clear differences between the diachronic samples in their >2cM LSAI sharing (Table S4). Further, we observe that the sharing of >5cM and >7cM LSAI is at comparable levels for modern-modern and modern-medieval pairs of samples while >10cM segments can be detected more abundantly in modern-modern than modern-ancient pairs, likely because of the excess of distant genealogical relationships among the modern samples that would be absent in modern-ancient pairs.

To explore further regional details of LSAI sharing patterns between Estonians and their geographic neighbours in light of evidence from ancient imputed genomes we focused on $PiC_{L>5cM,k>0.0005}$ 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 [62] 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, Belorussians, Lithuanians) (Figure 3) with whom they share Steppe ancestry [63] 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 year time depth show >10x 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) Belorussians 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 fixed yet in the Iron Age.

In contrast to individuals sampled from earlier periods, six Estonian medieval genomes from 13th-16th cc share significantly more >5 cM LSAI 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 LSAI than 82 EstBB Latvians and Lithuanians (average 2%, range 0-7%) or a medieval low coverage genome from Cambridgeshire, UK, imputed with Estonian medieval samples (Figure 4). Medieval Estonians share, however, significantly more >5 cM LSAI with Finns (10.1% average, Table S3) than Iron Age (average 4.7%, $p=0.006$, 2-tailed t-test) or modern Estonians (8.7% on average) suggesting that recent (17^{th}–18^{th} cc) 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} cc as 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 Supplementary Methods). To summarize the compound effect of imputation errors on LSAI-based ancestry mapping of the ancient samples we applied the UMAP dimension reduction method on the regional (county-based) PiC scores (Figure 5) and observed that the imputed TPM003i0.1x mapped closely to its high coverage version that had not been imputed. This, along with regional clustering of other medieval genomes among Estonian Biobank individuals from the same geographic context suggests that (i) LSAI inference through imputation from ancient low coverage genomes can be achieved sufficiently accurately for addressing questions about regional ancestry, (ii) >5cM LSAIs 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, (iii) 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 12th/13th cc AD from 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 using any geographic or other sample pooling criteria. We extracted communities 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 and twenty 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 3rd-level sub-clusters I7b of a 2nd-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 [56] in order to better understand the observed patterns of extensive LSAI sharing between ancient and modern Estonian genomes and how these are affected by demographic history (Figures S5-6). The simulation models were inspired by effective population size (Ne) trajectories obtained by applying IBDNe [64] to modern high-coverage Estonian genomes [33]. 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 an average individual is connected to is inversely and linearly dependent on Ne (compare Figure S6 A/C vs F/H) 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 with the specific pattern being 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, with a notable drop deeper in time (Figure S5 A/C, B/D and E/G). 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 ca 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 S5 F/H, 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 [46]. 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 (CI 95% 2300-1800) 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 (Burger et al. 2020) 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 which were sufficiently well covered in the high coverage TPM003 genome (Table S13).

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 [65]. rs1426654 is a variant that explains major part of skin pigmentation differences between Africans and Europeans, and differences among South Asians [66-68]. 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 [69]. 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 minor allele frequency is reduced and potentially problematic in selection targeted regions of the genome [70], we further assessed the accuracy of our LSAI inference from imputed data by comparing TPM003-s LSAI sharing, using a >2cM 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 5cM, there are both G- and A-allele carrying haplotypes shared between TPM003 and 4 Estonians (3 with the A allele and 1 with the G allele) and 1 Finn (with the G allele) (Figure S7, panel A). On average, G-allele carrying haplotypes are significantly longer (3.44 cM vs. 2.42 cM, one-tail t-test: 6.63E-09), 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 [65]. The core 200kb 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 JPT individual from 1000 GP) (Figure S7, panel B).

Discussion

We have shown that shotgun sequencing of ancient DNA at low (0.1-1x) 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 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; 71]. The increased accuracy can potentially be explained by (a) the temporal (and genetic ancestry) proximity of our medieval genome compared to the Neolithic sample considered in Hui et al. (2020) and (b) 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 notable drop in imputation accuracy we caution against generalizations to cases without ethnically/regionally matched large reference panels, 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 (0.1x) 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 0.3x 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 [72] or GRUPS [73] offer accurate estimation of close relatives from low (>0.05x) 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 to infer matches in genealogical relationships (up to 5th cousins). However, our simulations (Figures S5-6) 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 4-5th 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 [34], 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 4cM 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 [62]. 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 [33] (Figure 1F) by historically attested Finnish settlements in Northeast Estonia in the 17th–18th cc. However, our ancient DNA evidence (Figures 3-4) from 12th-16th cc 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; 20; 74] 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) not only since the Iron but since the Bronze Age (Figures 3-4) 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, with individuals from the 12th–14th cc AD showing 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 cc 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 data sets (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 [31-33] further research would be required to determine 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 due to 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 modelling of Finnish effective population size history in context of its local regional diversity and admixture sources.

Supplemental Information description

Supplemental material including six figures and fifteen tables can be found online.

Acknowledgements

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 using 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.), 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.), 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 by European Regional Development Fund 2014–2020.4.01.16–0030 (C.L.S.) and K.T. by UT Institute of Genomics grant PP1GI19936. This research has been conducted using the UK Biobank Resource under Application Numbers 54698 and 19947.

Declaration of interests

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

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/data/ (accession code PRJEB46155) and the data depository of the EBC (http://evolbio.ut.ee) .

The Estonian Biobank (EstBB) data used in this study are available under restricted access. The procedure of applying for the access to the data can be found under the following link: https://genomics.ut.ee/en/biobank.ee/data-access.

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 co-ancestry connections between the genome of a Medieval individual and modern Europeans. Sci Rep-Uk 10.
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. bioRxiv.
7. Cassidy, L.M., Maolduin, 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., Domboroczki, L., Kovari, 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., O'Maolduin, 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-Uk 10.
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.M., Tang, K.C., Zhang, S.J., and Zhi, D.G. (2019). RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. Genome Biol 20.
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. bioRxiv.
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., et al. (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 e1716.
20. Tambets, K., Yunusbayev, B., Hudjashov, G., Ilumae, 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.
21. Lang, V. (2020). Homo Fennicus.(Helsinki: Suomalaisen kirjallisuuden seura).
22. Honkola, T., Vesakoski, O., Korhonen, K., Lehtinen, J., Syrjanen, K., and Wahlberg, N. (2013). Cultural and climatic changes shape the evolutionary history of the Uralic languages. J Evolution 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. 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. 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. American Journal of Human Genetics 102, 760-775.
32. 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-Genes Genom Genet 7, 3459-3468.
33. 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.
34. Freyman, W.A., McManus, K.F., Shringarpure, S.S., Jewett, E.M., Bryc, K., Me Research, T., and Auton, A. (2021). Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows-Wheeler Transform. Molecular biology and evolution 38, 2131-2151.
35. Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., et al. (2015). A global reference for human genetic variation. Nature 526, 68-+.
36. 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.

37. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc
38. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17, 10-12.
39. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
40. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
41. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.
42. Jonsson, 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.
43. 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.
44. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15, 356.
45. 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.
46. 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.
47. Abraham, G., Qiu, Y.X., and Inouye, M. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics 33, 2776-2778.
48. Genomes Project, C., 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., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.
49. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. Plos Genet 2, 2074-2093.
50. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Use R, 1-212.
51. Karmin, M., Saag, L., Vicente, M., Wilson Sayres, M.A., Jarve, M., Talas, U.G., Rootsi, S., Ilumae, A.M., Magi, 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.
52. 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. (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet 48, 593-599.
53. Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics 47, 1-34.
54. Speed, D., and Balding, D.J. (2015). Relatedness in the post-genomic era: is it still useful? Nat Rev Genet 16, 33-44.
55. Foundation, F.S. (2014). GNU Datamash. In Retrieved from https://wwwgnuorg/software/datamash/. (

56. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. Plos Comput Biol 12.
57. Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems 1695, 1-9.
58. Hothorn, T., and Hornik, K. (2019). exactRankTests: Exact Distributions for Rank and Permutation Tests. In R package version 08-31. (
59. 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. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.
60. 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.
61. 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.
62. 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.
63. 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 e2186.
64. Browning, S.R., and Browning, B.L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. American Journal of Human Genetics 97, 404-418.
65. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X.H., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913-U912.
66. 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.Y., Humphreville, V.R., Humbert, J.E., et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310, 1782-1786.
67. Mallick, C.B., Iliescu, F.M., Mols, M., Hill, S., Tamang, R., Chaubey, G., Goto, R., Ho, S.Y.W., Romero, I.G., Crivellaro, F., et al. (2013). The Light Skin Allele of SLC24A5 in South Asians and Europeans Shares Identity by Descent. Plos Genet 9.
68. Norton, H.L., Kittles, R.A., Parra, E., McKeigue, P., Mao, X.Y., 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. Molecular biology and evolution 24, 710-722.
69. 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.
70. Burger, J., Link, V., Blocher, J., Schulz, A., Sell, C., Pochon, Z., Diekmann, Y., Zegarac, A., Hofmanova, 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-+.
71. Davies, R.W., Flint, J., Myers, S., and Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. Nat Genet 48, 965-969.
72. Kuhn, J.M.M., Jakobsson, M., and Gunther, T. (2018). Estimating genetic kin relationships in prehistoric populations. Plos One 13.

73. Martin, M.D., Jay, F., Castellano, S., and Slatkin, M. (2017). Determination of genetic relatedness from low-coverage human genome sequences using pedigree simulations. Molecular ecology 26, 4145-4157.
74. 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.
75. Skoglund, P., Malmstrom, H., Omrak, A., Raghavan, M., Valdiosera, C., Gunther, T., Hall, P., Tambets, K., Parik, J., Sjogren, K.G., et al. (2014). Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. Science 344, 747-750.
76. Skoglund, P., Malmstrom, H., Raghavan, M., Stora, J., Hall, P., Willerslev, E., Gilbert, M.T., Gotherstrom, A., and Jakobsson, M. (2012). Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336, 466-469.
77. 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.

Figures

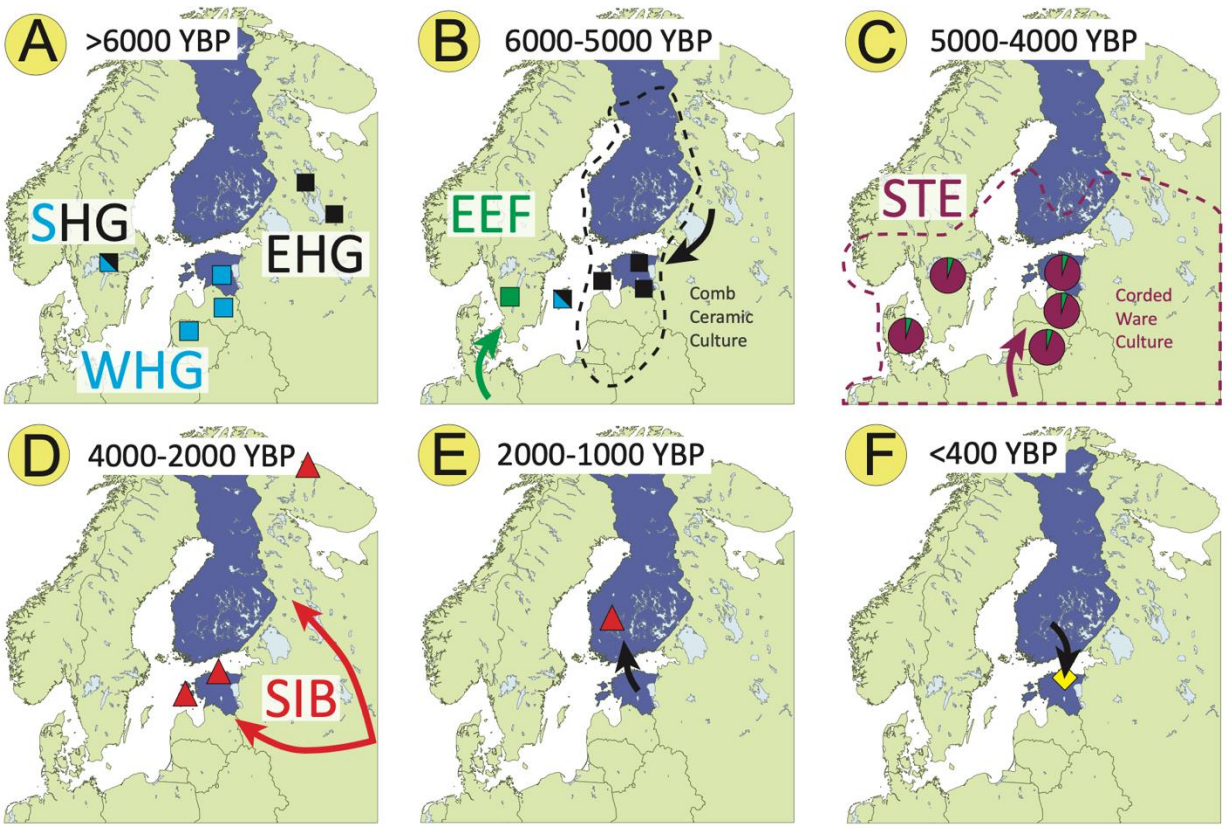
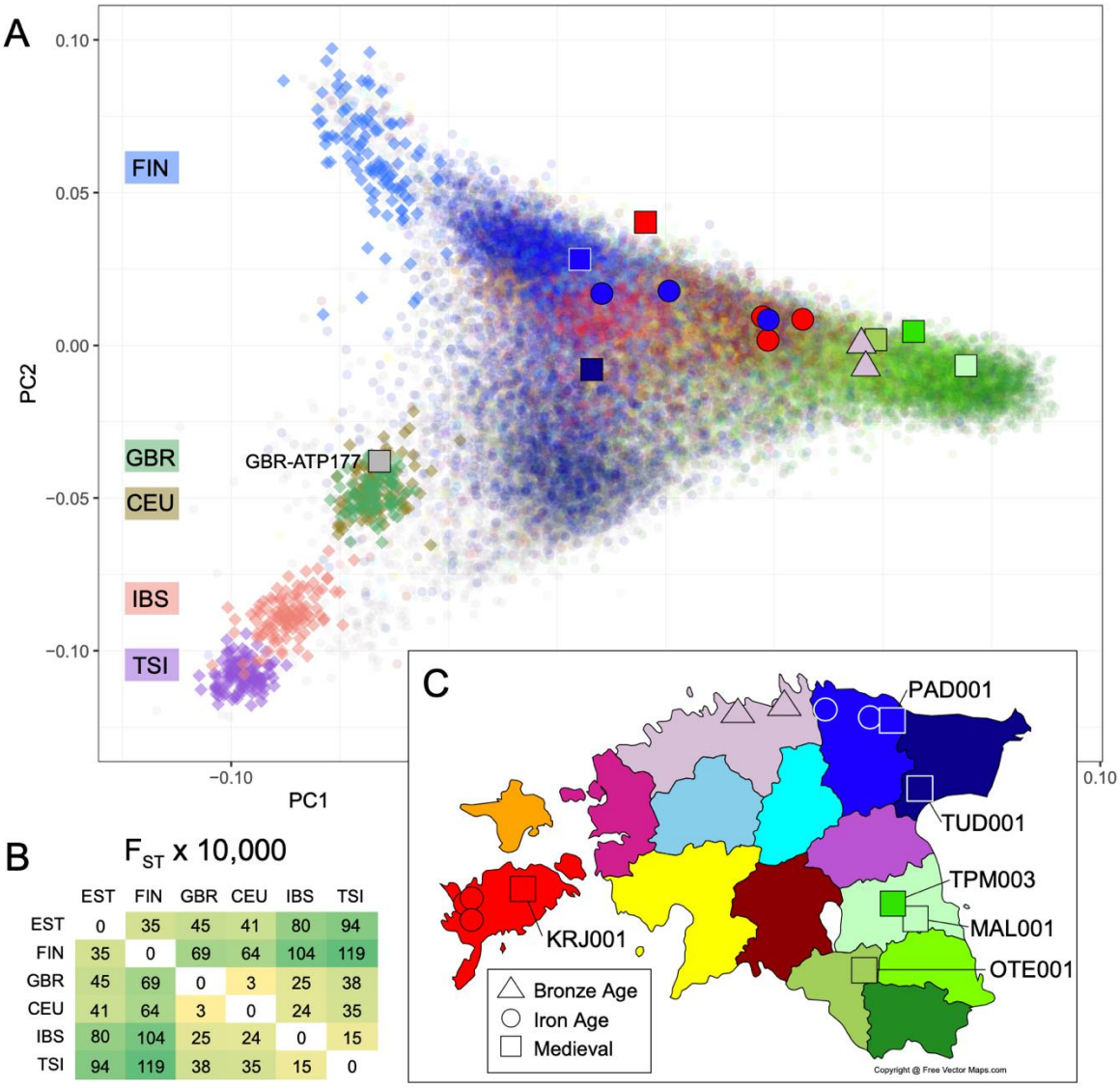


Figure 1 Simplified summary of population history of Estonia and Finland. **A-C.** Initial settlement by EHG - Eastern Hunter-Gatherers, WHG - Western Hunter-Gatherers, SHG - Scandinavian Hunter-Gatherers (the locations of sites that have produced ancient DNA evidence are shown with squares), followed by inflow of EEF - Early European Farmer and STE - Steppe ancestry [19; 63; 74-77] (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 2000–1000 years before present (YBP), with Honkola et al. [22] analyses placing the

split time at 800 AD. **F.** Some Finnish settlements in Northeast Estonia (highlighted with a yellow diamond) date back to 17th-18th cc.



	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 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 using projection. **B.** Fst estimates among modern populations

where EST refers to Estonians from the Estonian Biobank 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.

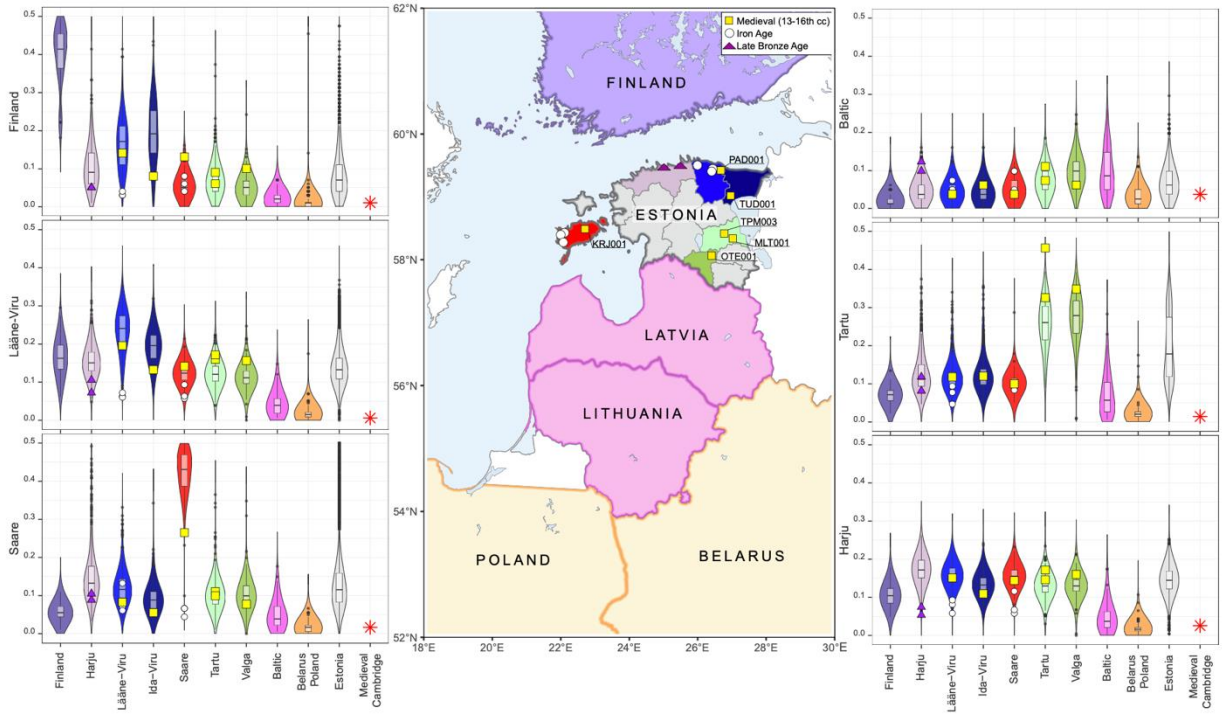


Figure 3 Genetic connectedness with six target populations of the circum-Baltic region and Estonian Bronze, Iron Age and medieval genomes. Each individual violin plot shows the distribution of PiC scores which reflect individual probabilities of >5 cM LSAI sharing and kinship coefficient estimate higher than 0.0005 with individuals from the modern target 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 from the 1000 GP data, 4739 EstBB Estonians born before 1940 including 1880 from 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 PSN177 - as a control, using 2,092 Estonian high coverage sequences as a reference panel.

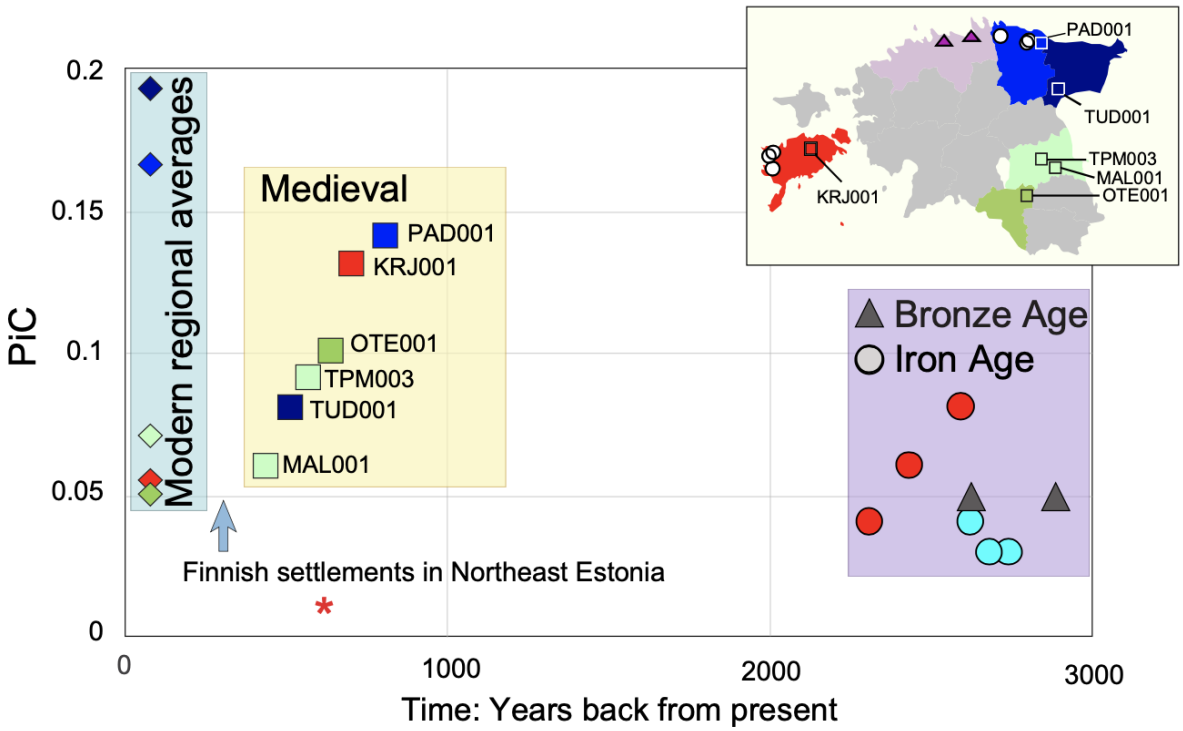


Figure 4. Temporal changes of individual connectedness with Finns in Estonian genomes from Bronze Age to present-day. Medieval, Bronze and Iron Age individuals are each represented by a square, triangle or a circle, respectively. Individual connectedness with Finns (PiC_{FIN}) of individuals born before 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. Note, PiC - proportion of 1000 GP Finnish individuals with whom modern and ancient Estonians share at least one LSAI segment >5cM and kinship coefficient >0.0005; * - medieval British low coverage (0.1x) genome, PSN177, imputed together with medieval Estonians using reference panel of 2,092 Estonian high coverage whole genome sequences.

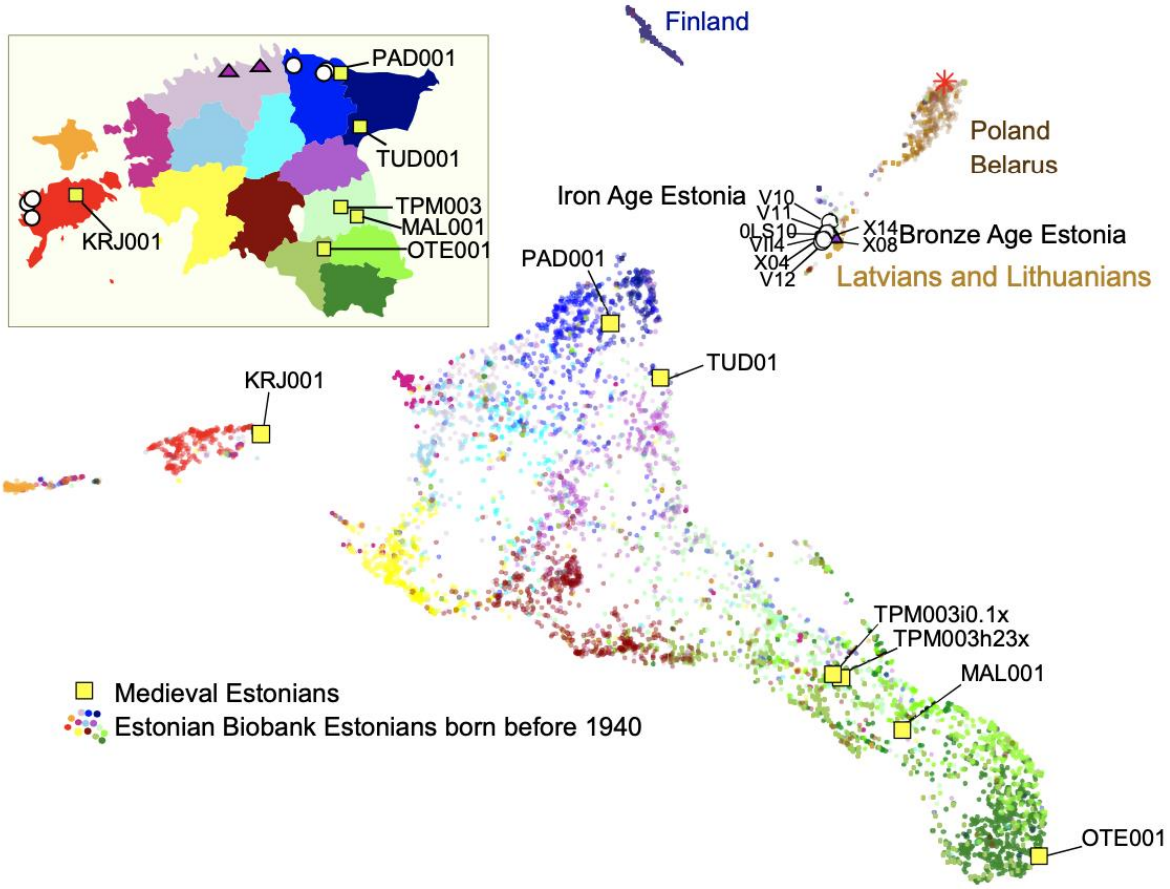


Figure 5. UMAP plot of LSAI sharing among medieval, Bronze and Iron Age genomes from Estonia, EstBB Estonians born before 1940, 1000 GP Finns, and EstBB Latvian, Lithuanian, Polish and Belarus samples. UMAP analyses were performed on 18 PiC score vectors estimated with L>5cM and k>0.0005 thresholds for 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 TPM003h23x were imputed. Six medieval Estonian genomes were imputed together with a medieval British low coverage (0.1x) genome PSN177 shown with red asterisk (*).

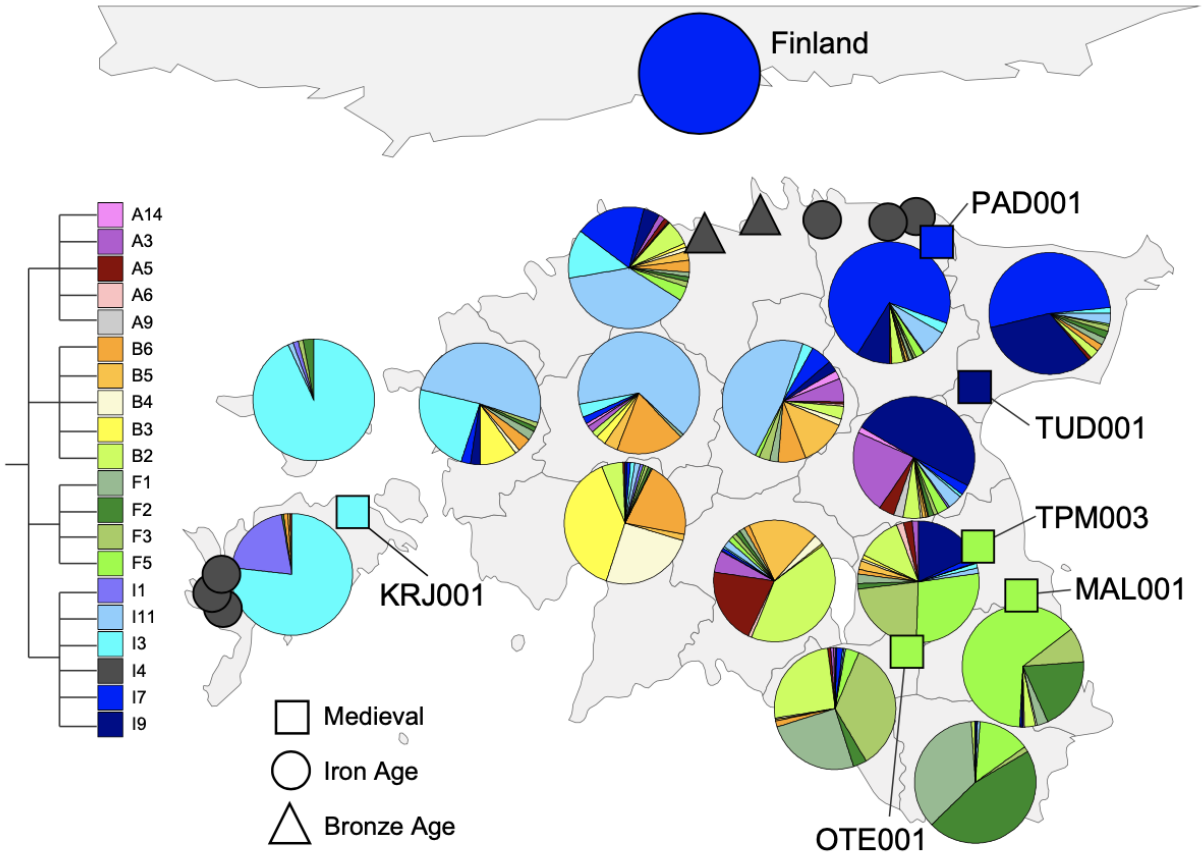


Figure 6. Two level community-structure inferred from 4739 Estonian Biobank individuals born before 1940, Finns of the 1000 GP and 14 ancient Estonians. Unsupervised community extraction method (Louvain method, [17]) was applied on 5cM LSAI-sharing signals estimated with IBIS. Four communities with more than 10 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 6 significant sub-communities (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). 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.