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

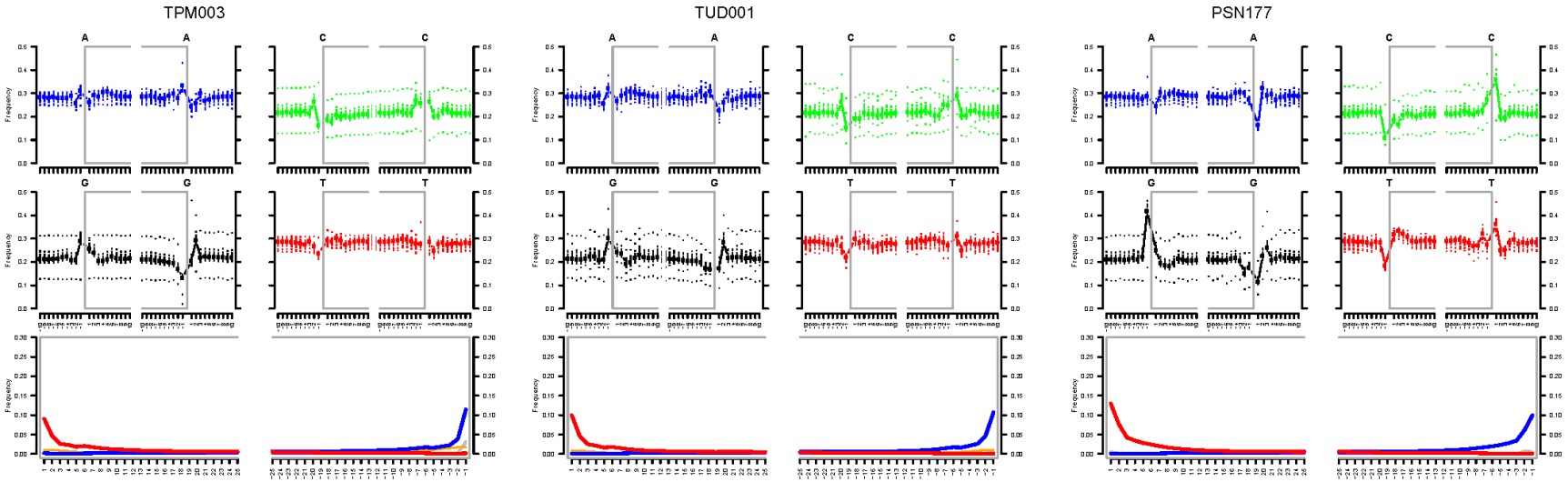


Figure S1. Genome-wide average damage rates in the three newly reported ancient genomes. The top plots show the frequency of each nucleotide in the 10 positions before and after each end of the DNA fragments that mapped to the human genome. The plots on the bottom show nucleotide substitutions at the 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.

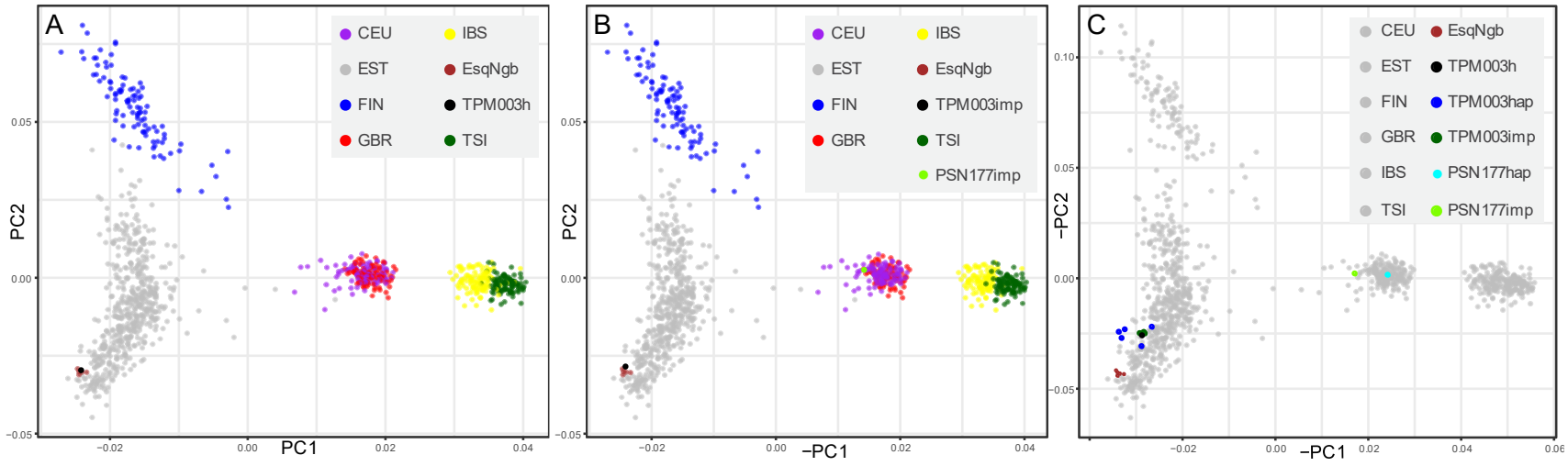


Figure S2. Smartpca analyses of TPM003 in context of 526 EstBB Estonians and 503 Europeans from 1000 GP. Genotypes were first directly called from the 23x coverage sample (TPM003h) without imputation. **A.** The placement of TPM003h on the PC plot is shown without projection. **B.** The placement on 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.

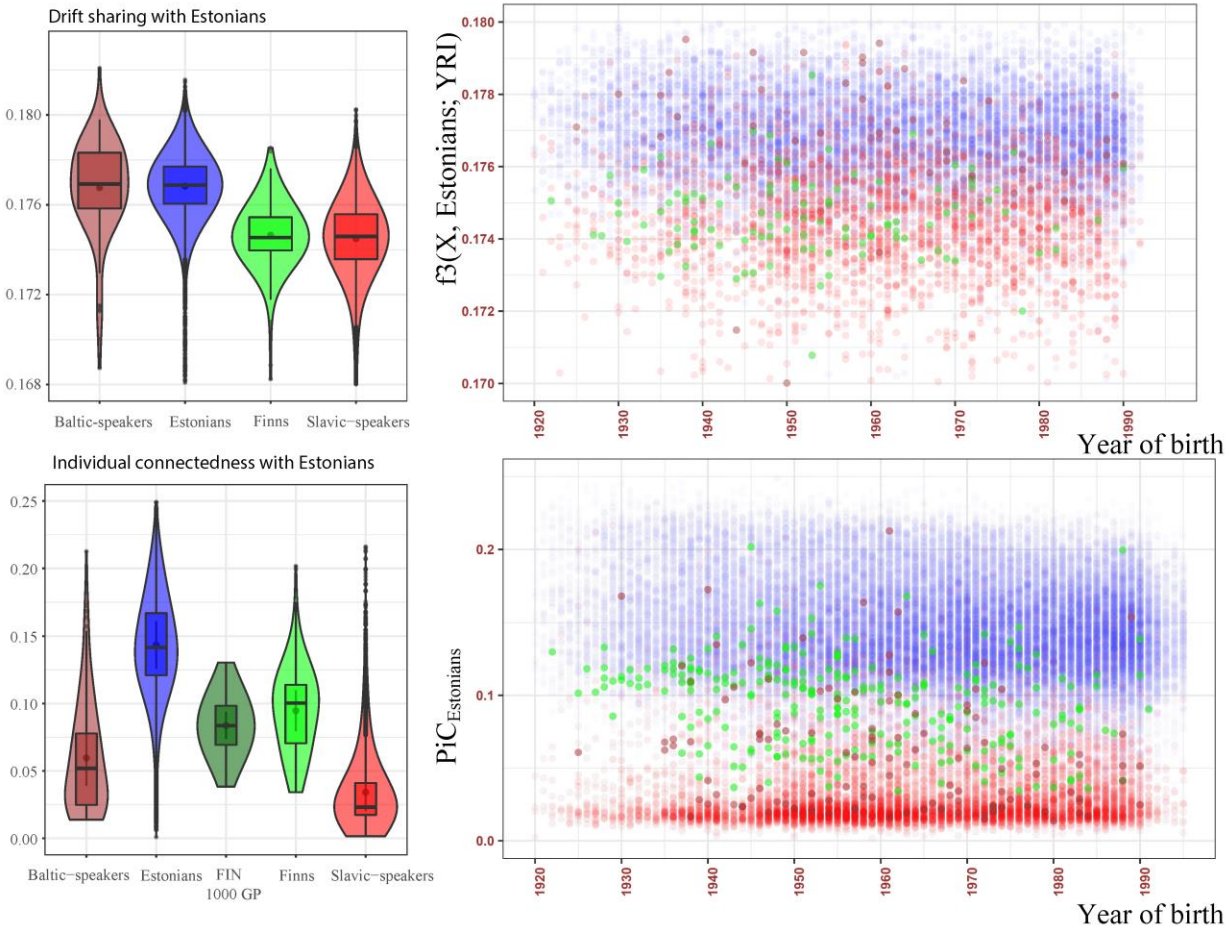


Figure S3. Drift sharing and individual connectedness with present-day Estonians. Upper panel shows drift sharing assessed using outgroup f3 statistics on Estonian, Finnish, Baltic and Slavic speaking subset of 47,015 EstBB individuals whose birthplace and year of birth data was available. Lower panel 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).

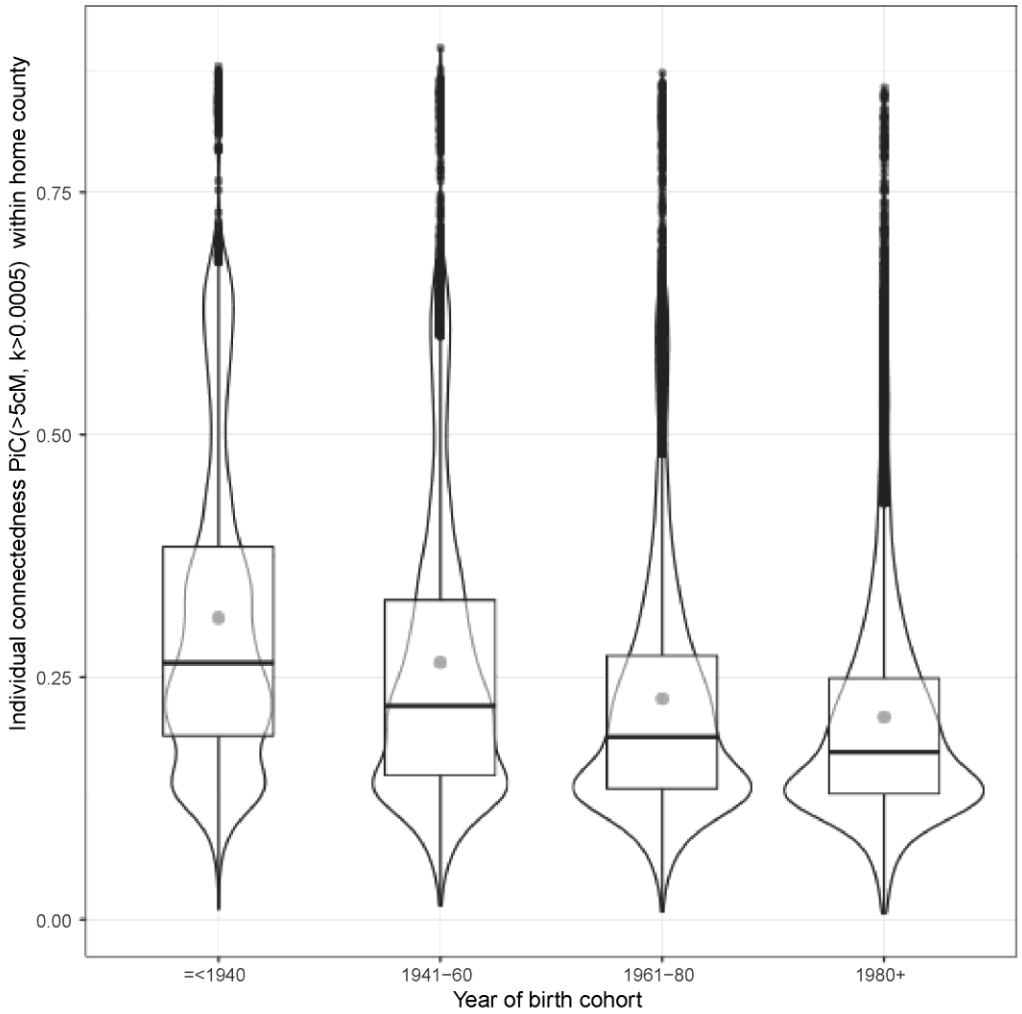


Figure S4. Proportions of individual connectedness of 47,015 EstBB samples (with birth place and year of birth information) within their home county by year of birth cohorts: (1) =<1940 cohort with 4,370 individuals, (2) 1941-60 cohort with 11,869 individuals, (3) 1961-80 cohort with 18,101 individuals,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

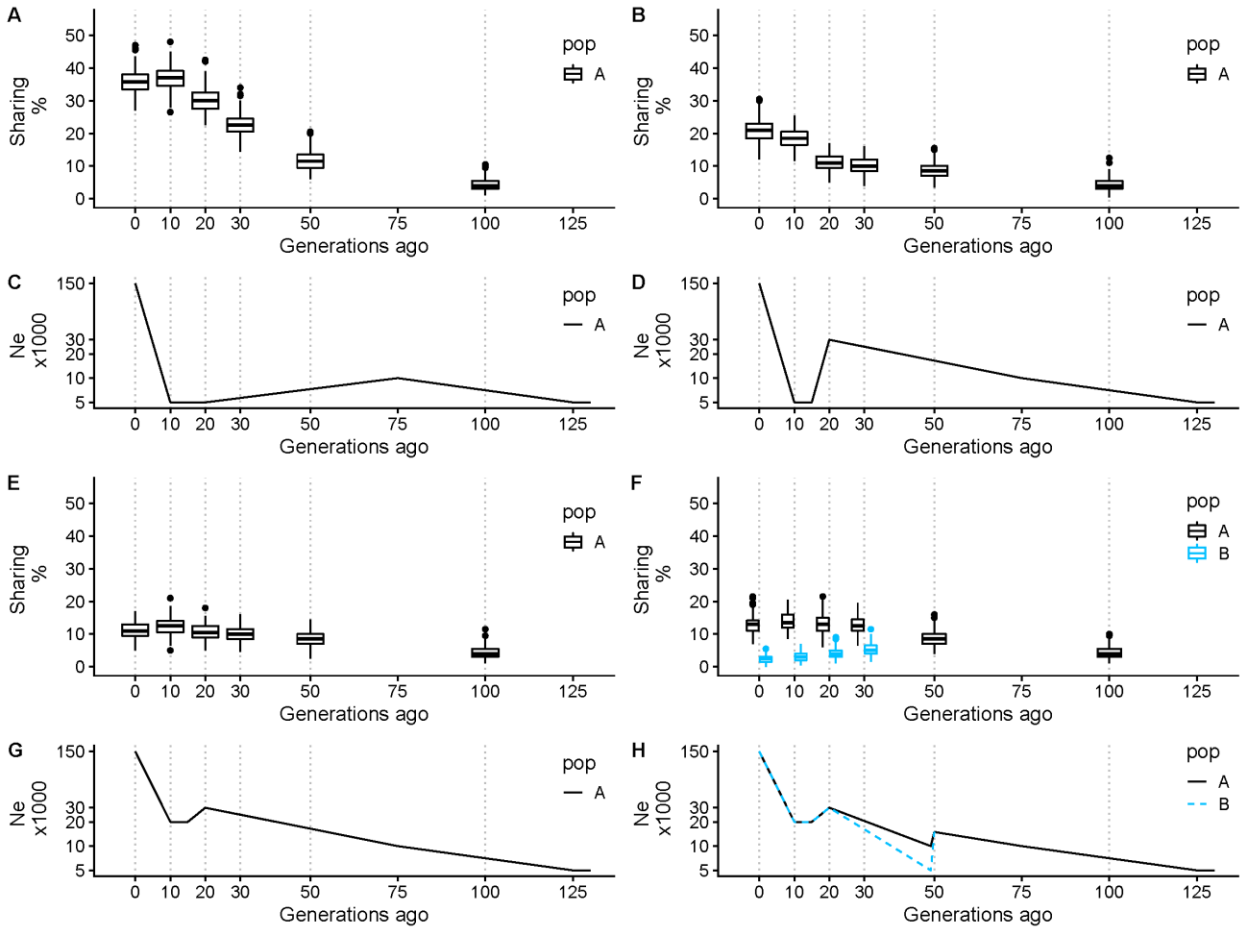


Figure S5. Individual connectedness among modern and ancient samples simulated under demographic scenarios inferred from whole genome sequences of present day Estonians (Pakratov et al. 2020). Panels A, B, E and F show the percentage of samples (PiC score) from different time points sharing at least 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.

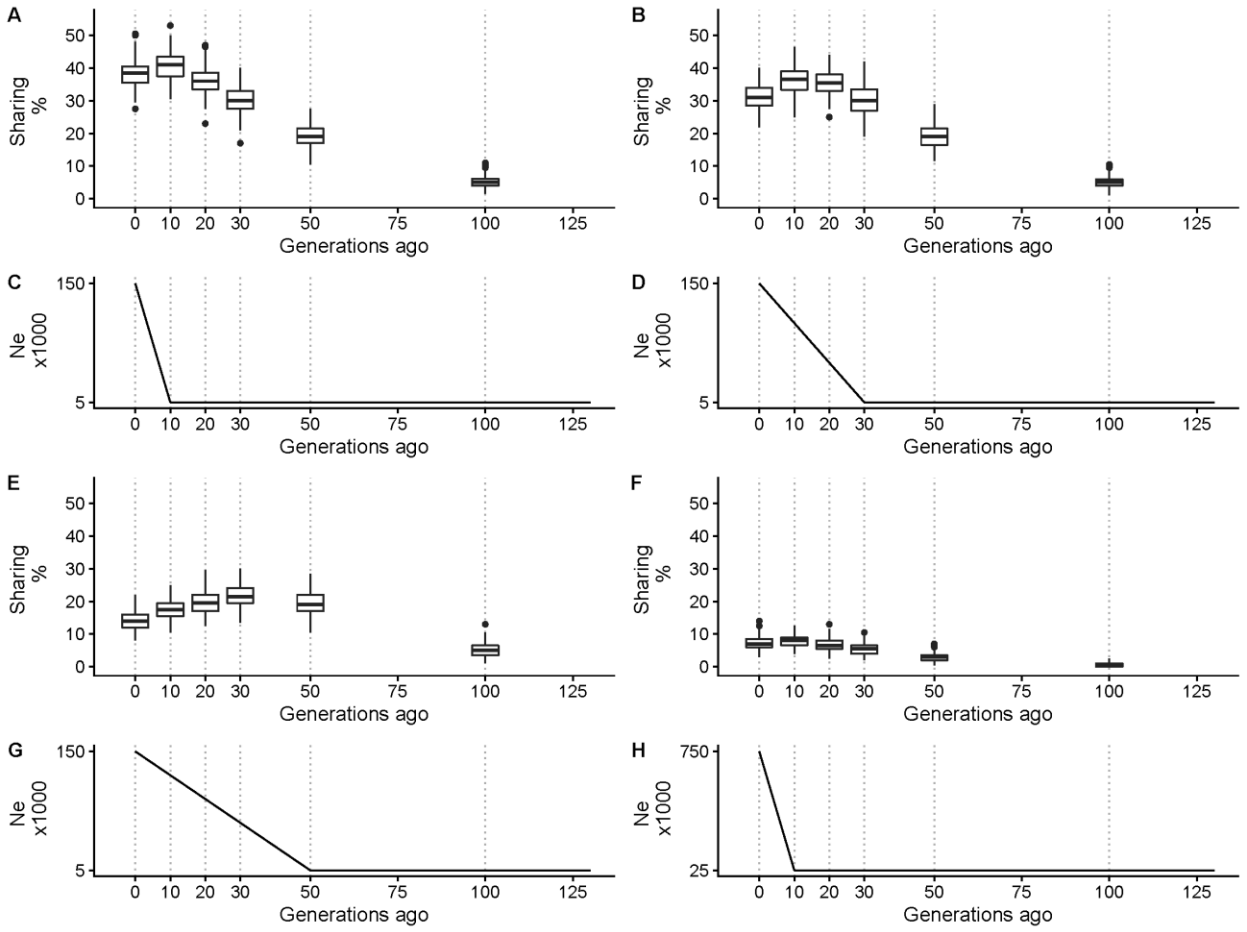
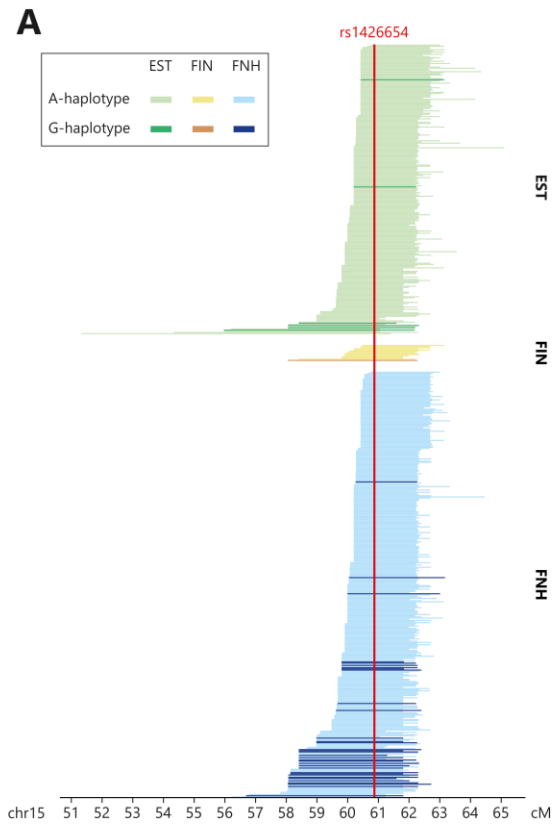


Figure S6. Individual connectedness among modern and ancient samples simulated under various demographic growth scenarios. Panels A, B, E and F show the percentage of samples (PiC score) from different time points sharing at least one IBD segment of 5 cM or longer and having a kinship coefficient above 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).

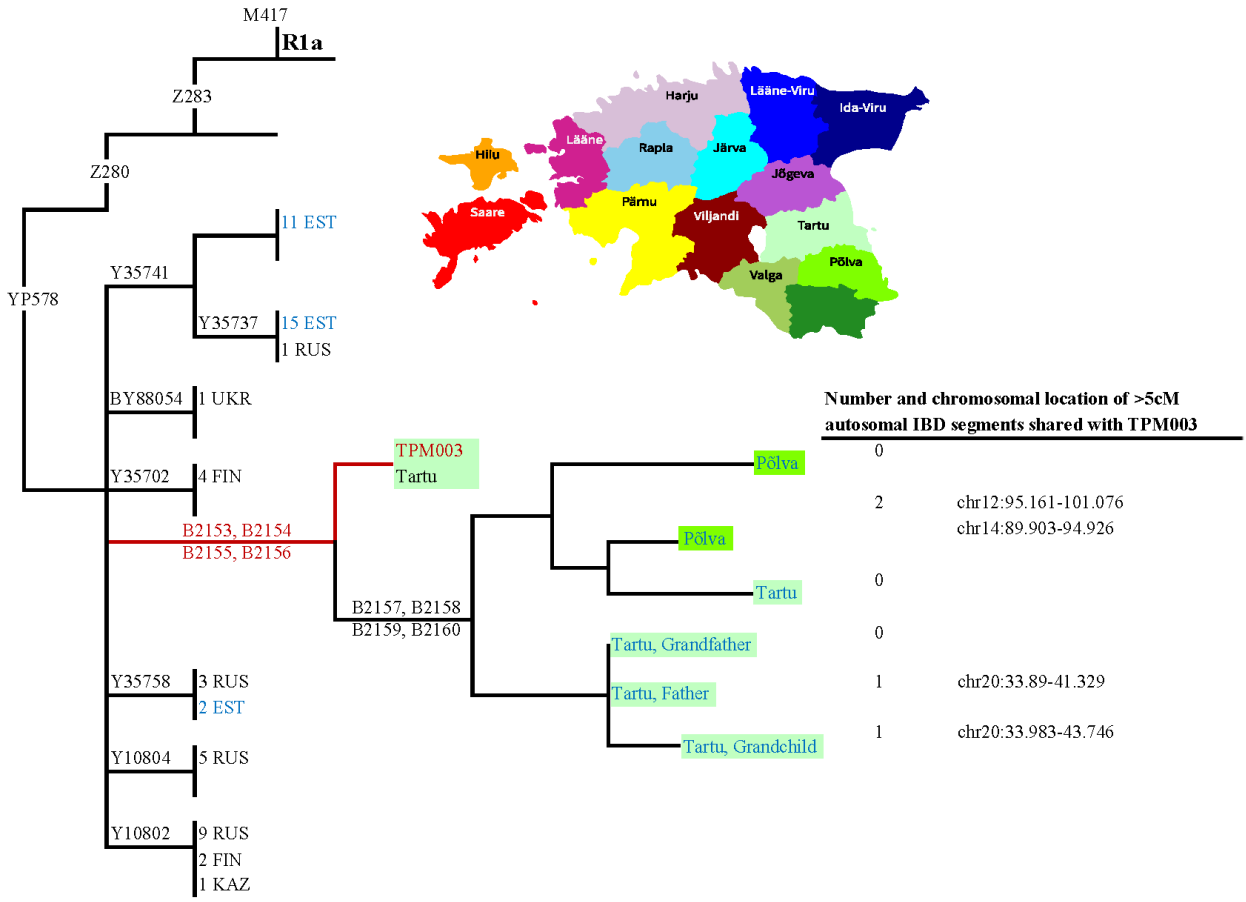


Figure S8. The placement of the medieval sample TPM003 in the context of phylogenetic tree of Y chromosome haplogroup R1a-YP578 in modern samples and his autosomal IBD sharing with present-day Estonians. Y chromosome lineages reported in the Y-Full YTree are shown in black and individuals from the Estonian 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.

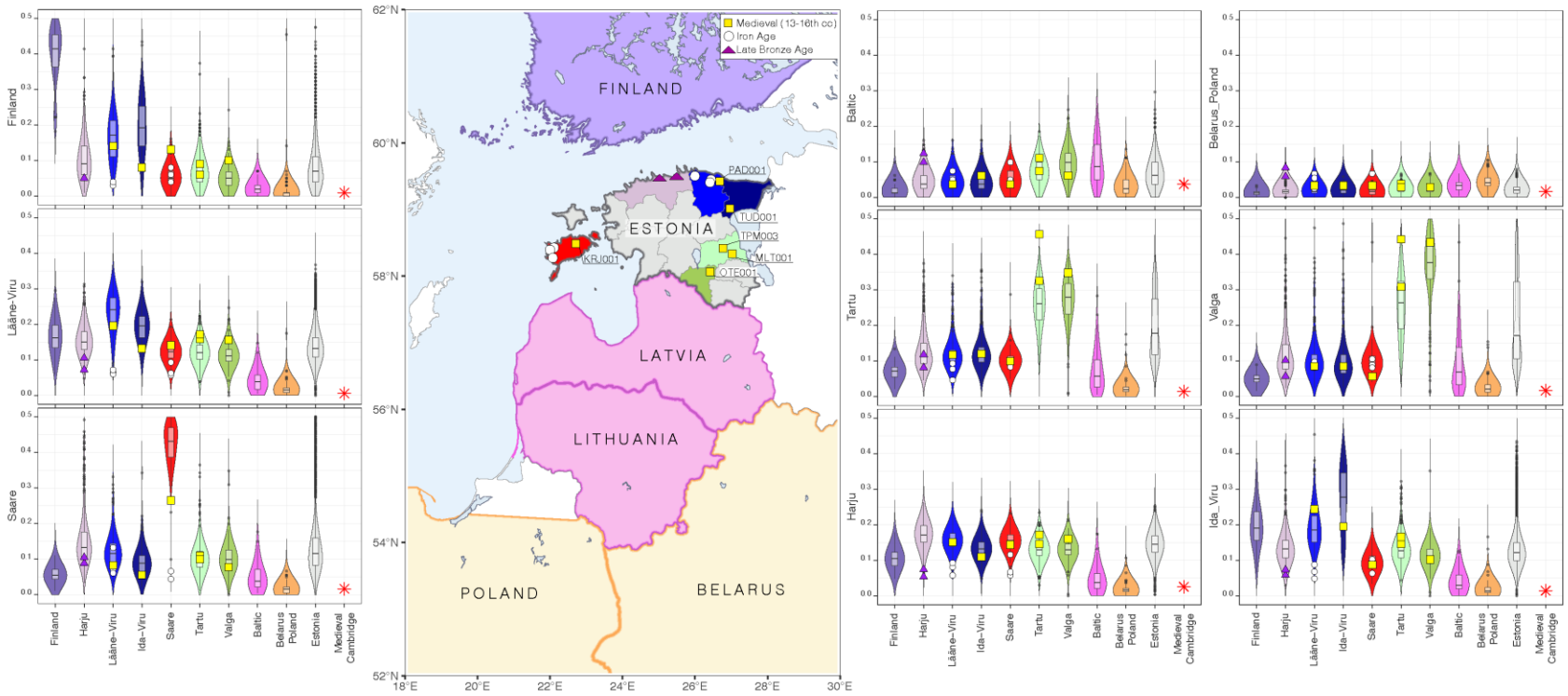


Figure S9. Genetic connectedness with nine target populations in modern populations of the circum-Baltic region and Estonian Bronze, Iron Age and medieval genomes. Each individual violin plot shows the distribution of PiC scores which reflect individual probabilities of >5 cM LSAI sharing and kinship 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.