Chapter 5

The genomics of the Slavic migration period, Early Middle Ages and their links to the present day

5.1 Introduction

The Slavic peoples originated as a diverse network of tribal societies who lived in Central and Eastern Europe from the first Millennia AD [159] and whose origin, although disputed, is thought to be Polesia (a marshy forested area straddling Poland, Belarus, Russiana and Ukraine) [160]. Although various Roman and Greek sources refer to Slavs as *Veneti* and *Spori* as early as the 1st and 2nd centuries AD, the term 'Slavs' was first used in writing by Roman bureaucrat Jordanes at the beginning of the 6th century after their attack on the Byzantine empire [161]. This era, known by historians as The Migration Period, was a period of European history, roughly between 375-568 AD after the fall of the Roman Empire [162], characterised by the large-scale movement of various peoples. The Migration Period began with the Huns moving into Eastern Europe at the end of the 4th Century, occupying an area including present-day Hungary and Romania. During the 5th century, various Germanic groups invaded and established a homeland across parts of the Western Roman Empire. This was followed by the expansion of Slavic populations into regions of low population density in the sixth century.

Across the next 2 centuries, these peoples had settled across large parts of Europe. In particular, the Early Slavs had expanded southwards into the Balkans and Alps [159,163–165]. It has been proposed that these migrations were key to forming the foundations of present-day Slavic (speaking) nations [159].

By the beginning of the 12th century, Slavs constituted a large part of a number of many medieval Christian states across Europe. As from this time period, Slavs could be broadly split up in 3 groups: the Eastern Slavs as part of the Kievan Rus', Southern Slavs in the Bulgarian Empire, the Principality of Serbia, Kingdom of Croatia and the Banate of Bosnia, and Western Slavs in the Principality of Nitra, Great Moravia, the duchy of Bohemia and the Kingdom and Poland. In addition, Slavic settlement also occurred in the Eastern Alps; Slovenia, large parts of present-day Austria and Friul.

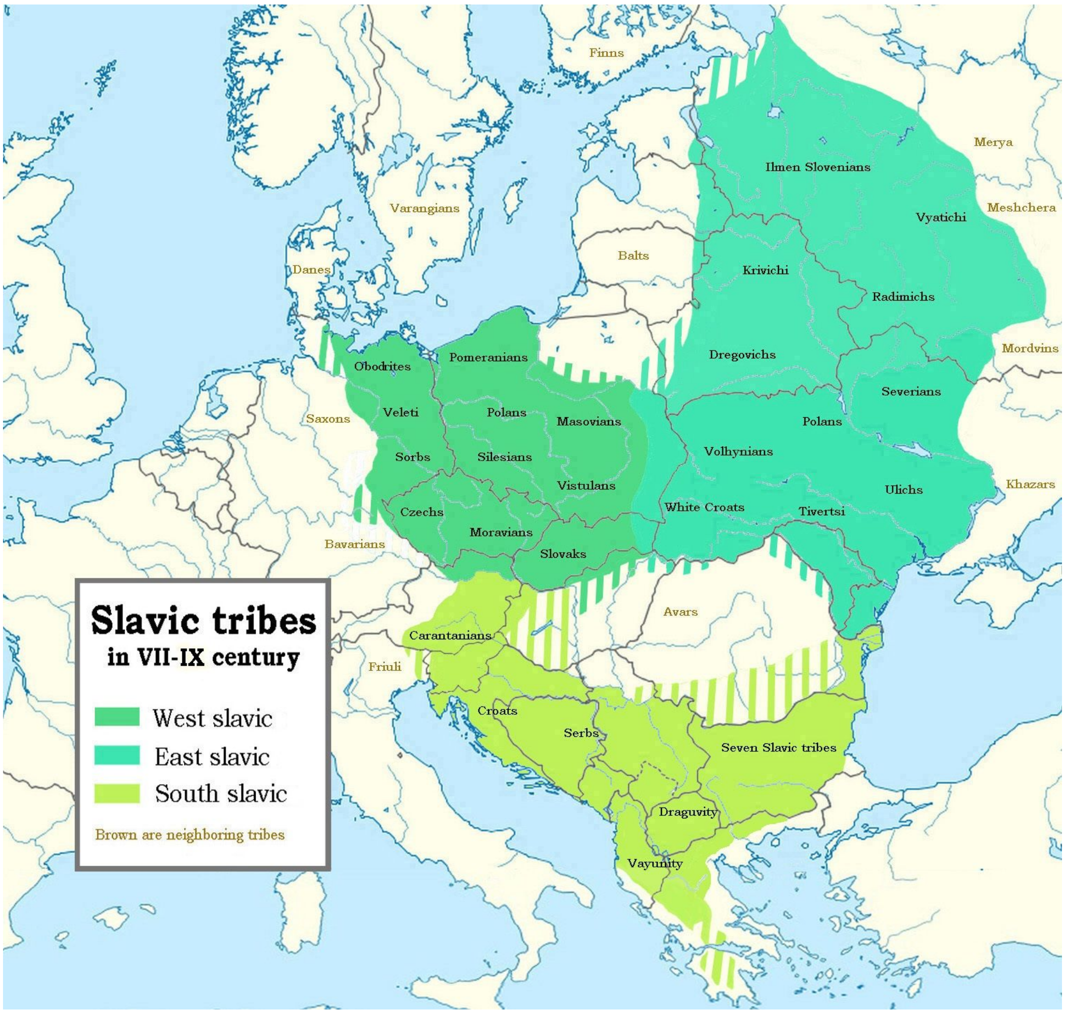

Figure 5.1: Slavic tribes from the 7th to 9th centuries AD in Europe. Source: wikipedia.

The differentiation of Slavs into these 3 broad groups can still be seen today in the different language groups. Today 315 million people speak Slavic languages. Linguistic evidence suggests that they can be broadly split into 3 groups; Western Slavs (Poles, Czechs and Slovaks), Eastern Slavs (Ukrainians, Belarusians and Russians) and Southern Slavs (Croatians, Bulgarians, Slovenians, Bosnians, Macedonians, Montenegrins and Serbians) [166].

The history of the Slavic peoples can be artificially be split into 3 periods; Migration

Period (3̃75AD - 568AD), Early Middle Ages/High Middle Ages (∼∼600AD - ∼∼1250AD) and present-day. Several previous studies have investigated the genetics of the transitions between these periods. Juras et al (2014) used uniparental mtDNA markers from ancient DNA samples from Poland to show continuity between both Roman Iron Age period (200 BC – 500 AD) and Medieval Age (1000–1400AD) with present-day Poles, Czechs and Slovaks [167]. Whilst informative about sex-biased migrations, uniparental markers carry only a fraction of the information that autosomal markers do, and therefore may provide misleading or incomplete information about the relationship between samples [168, 169], especially when admixture is prevalent (although see [170]). For example, it is know that mtDNA and nuclear DNA may have different evolutionary histories and thus display discordant phylogenetic trees [171].

Kushniarevich et al (2015) [172] combined results from mtDNA, non-recombining Y and autosomal DNA to investigate the population structure of a wide range of present-day Balto-Slavic populations in order to understand the historical processes that have formed the present-day genetic structure. They proposed that admixture of incoming Slavic speakers during the Migration Period with the pre-existing substrate of regional genetic components, which differed between South, East and West Slavs. Using this evidence, they propose that the "cultural assimilation of indigenous populations by bearers of Slavic languages as a major mechanism of the spread of Slavic languages to the Balkan Peninsula".

More recently, Macháček et al (2021) [173] analysed ancient rune inscriptions on a cattle rib from Lány, Czechia, dated to approximately 600AD. The bone is inscribed with Germanic runes. Finding Germanic runes in the context of Slavic peoples provides evidence of early interactions between Slavic and Germanic peoples. The bone was found in a location where Slavs were thought to have arrived at the end of the Migration Period, after the Germanic tribes had disappeared and the use of a Slavic language is historically confirmed as of the 9th century. However, whether there was early genetic contact as well is yet to be determined.

Several studies into present-day Slavic populations have detected signatures of admixture from East-Asia [15, 144, 174–176]. Whether or not these signals can be observed in ancient individuals is yet to be seen and could further refine the admixture date. For example, different admixture dates in different Slavic populations may reveal structure among present-day Slavs.

Finally, several studies have used haplotype-based methods to explore the structure of present-day Slavic populations. Ralph and Coop [177] compared regions of IBD matching across different European populations. They found a relatively high degree of IBD sharing among pairs of individuals from Eastern Europe, suggestive of expansion from a smaller, common source population. This expansion was tentatively estimated to between 0-1000AD. Consistent with estimates of a small population size, Hellenthal et al (2014) [15] inferred an excess of IBD-sharing among Eastern European individuals, albeit with a more constrained admixture data of 440 - 1080 CE. However, this could also be interpreted in terms of a small effective population size [178, 179]. Salter-Townshend and Myers (2019) also identified admixture in the Chuvash people betweeen East Europeans and East Asians approximately 1224 CE [144].

Despite these efforts, no studies have integrated autosomal DNA from ancient, present-day samples and haplotype-based methods to infer population structure, ancestry proportions and admixture events. Therefore in this chapter, I will analyse 17 new medium to high coverage whole ancient genomes from Czech Republic, spanning the Migration Period and Early Middle Ages. These are, to my knowledge, the first high-coverage whole ancient-genomes from Slavic speakers. I will merge the newly sequenced samples with reference data from other ancient individuals and a large reference set of relevant present-day European individuals in order to understand their ancestry in the context of both present-day and ancient samples. In particular, I am interested in considering the following questions:

1. Can we gain an understanding of the geographical origins of the Slavic peoples from ancient DNA

2. Do the labels “Migration Period” and “Early Middle Ages” make sense from a genetic standpoint (i.e. do samples from either period cluster with another to the exclusion of the other)

3. Was there interactions between Germanic and Slavic peoples during the Early Migration Period.

4. If so, what genetic differences can be observed between these periods? Are they characterised by admixture from outside sources? If so, what are these sources and can the admixture events be dated?

5. What is the relationship between the ancient samples and present-day day Slavic populations. Are they continuous?

6. Do the different ancient Slavic samples have different affinities to different present-day Slavic language groups?

Code	Site	Date (AD)	Period	Coverage
LIB11	Břeclav – Líbivá	741.5	Early Middle Ages	5.3
LIB12	Břeclav – Líbivá	475.5	Migration period	6.8
LIB2	Břeclav – Líbivá	495.0	Migration period	6.4
LIB3	Břeclav – Líbivá	509.0	Migration period	5.3
LIB4	Břeclav – Líbivá	472.5	Migration period	6.5
LIB5	Břeclav – Líbivá	348.0	Migration period	7.3
LIB7	Břeclav – Líbivá	830.5	Early Middle Ages	5.6
POH11	Pohansko – Lesní školka	783.0	Early Middle Ages	5.0
POH13	Pohansko – Lesní školka	879.5	Early Middle Ages	6.0
POH27	Pohansko – Jizní Předhradí	783.0	Early Middle Ages	5.9
POH28	Pohansko – Jizní Předhradí	822.5	Early Middle Ages	5.6
POH36	Pohansko – Jizní Předhradí	880.5	Early Middle Ages	5.5
POH39	Pohansko – Jizní Předhradí	866.4	Early Middle Ages	5.3
POH3	Pohansko – Lesní hrúd	956.5	Early Middle Ages	5.4
POH40	Pohansko – Lesní školka	950.5	Early Middle Ages	5.5
POH41	Pohansko – Lesní školka	875.5	Early Middle Ages	5.2
POH44	Pohansko – Pohřebiště U Kostela	NA	Early Middle Ages	5.3

Table 5.1: Information on newly sequenced ancient samples. Date (AD) estimated from radiocarbon dating. 'Migration' corresponds to Migration Period and 'EMA' corresponds to Early Middle Ages. Coverage calculated as the mean depth across all 77,213,942 genome-wide SNPs where genotypes were called at.

5.2 Methods

5.2.1 Description of samples

Whole-genome sequence data was generated from 17 ancient individuals. All newly sequenced samples are from Czechia and are split across two different field sites.

The newly sequenced samples are grouped into two temporal categories; 5 samples are from the Migration Period (348 AD - 504 AD) and the Líbivá site, and the other 12 samples are from the later Early Middle Ages (724 AD - 995 AD) and the are from the Pohansko site.

Apart from the age of the samples, the Migration Period and Early Middle Age samples can be differentiated by the style of pottery found in the burial grounds (Z. Hofmanová, personal communication).

5.2.2 Ancient DNA processing

I merged the 17 newly sequenced individuals with the ancient literate samples given in section A.1 resulting in a total of 959 ancient individuals with genotype likelihoods at 77,213,942 genome-wide autosomal SNPs.

I followed the GLIMPSE [72] imputation and phasing pipeline (`https://odelaneau.github.io/GLIMPSE/tutorial_b38.html`) to generate genotype likelihoods and phased genotypes for each individual. For the reference panel, I used the 30x 1000 genomes dataset [83], described in appendix section A.2.

5.2.3 Present-day DNA processing

I chose to merge the newly sequenced and literature ancient samples with the MS-POBI-HellBus dataset, described in detail in appendix section A.4, because it contains a high number of relevant samples from central and Eastern Europe. I removed samples from Australia, New Zealand and USA, as these samples were not from native individuals from that country.

The present-day and ancient samples were phased separately. This was because GLIMPSE, which is necessary to phase the ancient samples with, is not suitable to phase the modern samples with, for two reasons. Firstly, GLIMPSE is designed to work with sequence-level density of data, and the modern samples have been genotyped on a low-density genotyping array. Secondly GLIMPSE accepts data as genotype likelihoods; these were not available for the modern samples. Therefore, the modern samples were phased using shapeit4 [20].

Once the present-day dataset was phased, I merged it with the phased haplotypes of the ancient samples described in secion 5.2.2 and converted to ChromoPainter format.

5.2.4 plink PCA

To determine the broad-scale ancestry distribution of the newly sequenced and ancient literature samples, I performed PCA on the pre-imputation genotypes using plink2 [142]. Performing an unlinked PCA also allows us to identify any data quality issues which are independent of phasing / haplotype-based analysis. I chose to use plink2 for its ease of use (no conversion to exotic format is required) and because recent studies have shown it is substantially better at dealing with samples containing variable amount of missing data than other methods such as smartPCA [49].

I retained only the 500,000 markers with the lowest amount of missingness and then LD-pruned the resulting SNPs using the settings `-maf 0.01` and `-indep-pairwise 50 5 0.2`. PCA was performed using default settings from plink2 and the first two principle components plotted.

5.2.5 Allele-frequency based tests

I used Admixtools [35], implemented in Admixr R library [180] to employ several different F-statistics. I chose to use F-statistics in addition to ChromoPainter analysis for two reasons. Firstly, it is possible to explicitly test models of population history (tests of treeness, admixture) in a more simple way than when using ChromoPainter. Secondly, F-statistics can be used on much lower coverage samples than ChromoPainter [37].

5.2.6 ChromoPainter and fineSTRUCTURE analysis

I began with a merged dataset of present-day and ancient individuals, described in sections 5.2.2 and 5.2.3 in ChromoPainter format. This dataset contained a total of 959 ancient and 14,795 present-day samples genotyped at 477,417 autosomal bi-allelic SNPs.

I first selected all ancient samples above 2x coverage and performed an 'all-v-all' painting where each haplotype was compared to all other haplotypes in turn. 2x was somewhat conservatively chosen as a conservative threshold to reduce coverage related bias (my work in Chapter 2 section 2.6.4 showed that 0.5x is a suitable coverage threshold) whilst still retaining a suitable number of individuals. This allows for the characterisation of the ancestry of the newly sequenced ancient samples in the context of other ancient individuals. It is also the painting that can be used to perform fineSTRUCTURE clustering and tree building on ancient samples. Hereafter referred to as 'ancient' painting.

I also performed an 'all-v-all' painting of a selected group of present-day individuals and the newly sequenced ancient individuals. The populations retained are given in table **??**. Hereafter referred to as 'present-day painting'.

Both the 'present-day' and 'ancient' paintings were merged separately using chromocombine-0.0.4 (`https://people.maths.bris.ac.uk/~madjl/finestructure-old/chromocombine.html`).

The fineSTRUCTURE [14] clustering and tree building algorithm was applied to the chunkcounts ChromoPainter output, for both the 'present-day' and 'ancient' paintings. It was first run in MCMC mode using 1,000,000 burn-in MCMC iterations and 2,000,000 main MCMC iterations. It was then run in tree-building mode (-m T) using 100,000 burn-in and 100,000 main iterations.

This algorithm assigns individuals to genetically homogeneous clusters, estimates the 'true' number of clusters and builds a dendrogram of genetic similarity. This is particularly useful when combining many samples from different studies, as is the case with the 'ancients'

Population	Number of Individuals
HB:tsi	98
HB:spanish	34
HB:german	30
HB:french	28
HB:greek	20
HB:croatian	19
HB:hungarian	19
HB:norwegian	18
HB:southitalian	18
HB:polish	17
HB:romanian	16
HB:mordovian	15
HB:cypriot	12
HB:northitalian	12
HB:lithuanian	10
HB:siciliane	10
HB:westsicilian	10
HB:tuscan	8
HB:irish	7
HB:scottish	6
HB:germanyaustria	4
HB:welsh	4

Table 5.2: Name of population and number of samples used in the present-day ChromoPainter analysis

painting; the population label identifiers used by different studies may not be consistent with one another. Therefore, we can use fineSTRUCTURE groupings as population labels rather than external group labels.

Tree figures, co-ancestry matrix figures and principle component plots were generated using the fineSTRUCTURE R library (`https://people.maths.bris.ac.uk/~madjl/finestructure/FinestructureRcode.zip`).

5.2.7 SOURCEFIND ancestry proportion analysis

I used SOURCEFIND [16] to infer the proportions of ancestry by which each target (e.g. ancient) individual is most related to a set of surrogate populations. Each of the 47 clusters of ancient samples inferred by fineSTRUCTURE was analysed in turn, using the other 46 clusters to act as surrogates.

Each cluster was run across 3 independent MCMC runs, using 50,000 burn-in iterations, 500,000 main iterations, thinned every 5 iterations. All 3 MCMC runs were then combined to form an MCMC list using the coda R libary [87] and `mcmc` function to jointly estimate ancestry proportions and empirical credible intervals for each target population.

5.2.8 MOSAIC admixture analysis

I inferred admixture events, dates and proportions using MOSAIC [144].

I performed 2 different kinds of admixture modeling. First, I performed an 'ancient analysis' using the 47 fineSTRUCTURE clusters of ancient samples to assign groups. I analysed each cluster in turn, using all other 46 clusters as surrogates.

I then performed a 'present-day surrogates' analysis using a select group of present-day populations **??** and all ancient Slavic samples. I analysed each population in turn using all other populations as surrogates.

MOSAIC was run using default settings and the following sets of populations as targets and the following sets as surrogates. I formed each target as a mixture of either 2 or 3 ancestral sources. Upper and lower quantiles for admixture dates were estimated using a bootstrap procedure.

5.3 Results

Principle Component Analysis (PCA) using plink2 showed that the Migration Period samples do not all cluster together and instead fall on a cline of similarity between a cluster of Central European Middle Age/Iron Age samples (top-left) and Neolithic samples (bottom-right) (Fig. 5.2). The Early Middle Age samples are more homogeneous, with all samples occupying the broad region containing European Iron Age samples. However, samples POH39 and POH3 display an elevated affinity to samples from Early Bronze Age Ireland.

5.3.1 Mixed ancestry of migration period Slavs

In order to reveal further structure in the ancient samples, I performed an all-v-all painting of 152 ancient samples with a coverage greater than 2x. I then applied the fineSTRUCTURE clustering algorithm to the samples in order to assign them to genetically homogeneous groups. Full details of each sample and their fineSTRUCTURE cluster given in Table E.1.

The Migration Period samples consisted of 5 individuals from Břeclav (Líbivá), Czech Republic, from 5 different burial sites, who had radiocarbon dates corresponding to the Migration Period (348 - 509AD). It is apparent from both the unlinked (Fig. 5.2) and linked PCAs (Fig. 5.3) that the Migration Period samples represent a heterogeneous group of individuals who do not originate from the same source population. LIB2 (495AD) is located in the centre of a large cluster of contemporaneous individuals from Iron Age Central and

Population	Number of Individuals
HB:han	34
HB:bulgarian	31
HB:japanese	28
HB:sardinian	28
HB:russian	25
HB:yakut	25
HB:greek	20
HB:ukrainian	20
HB:croatian	19
HB:hungarian	19
HB:mongolian	19
HB:southitalian	18
HB:chuvash	17
HB:polish	17
HB:romanian	16
HB:buryat	15
HB:mordovian	15
HB:altai	13
HB:tuva	13
HB:evenk	12
HB:northitalian	12
HB:cambodian	10
HB:dai	10
HB:hannchina	10
HB:lithuanian	10
HB:miao	10
HB:nganassan	10
HB:selkup	10
HB:siciliane	10
HB:tu	10
HB:tujia	10
HB:uygur	10
HB:westsicilian	10
HB:yi	10
HB:belorussian	9
HB:daur	9
HB:oroqen	9
HB:xibo	9
HB:hezhen	8
HB:naxi	8
HB:tuscan	8
HB:dolgan	7
HB:chukchi	5
HB:koryake	5
HB:yukagir	4
HB:myanmar	3
HB:burya	2
HB:ket	2
HB:malayan	1

Table 5.3: Name of populations and number of samples used in the present-day MOSAIC analysis

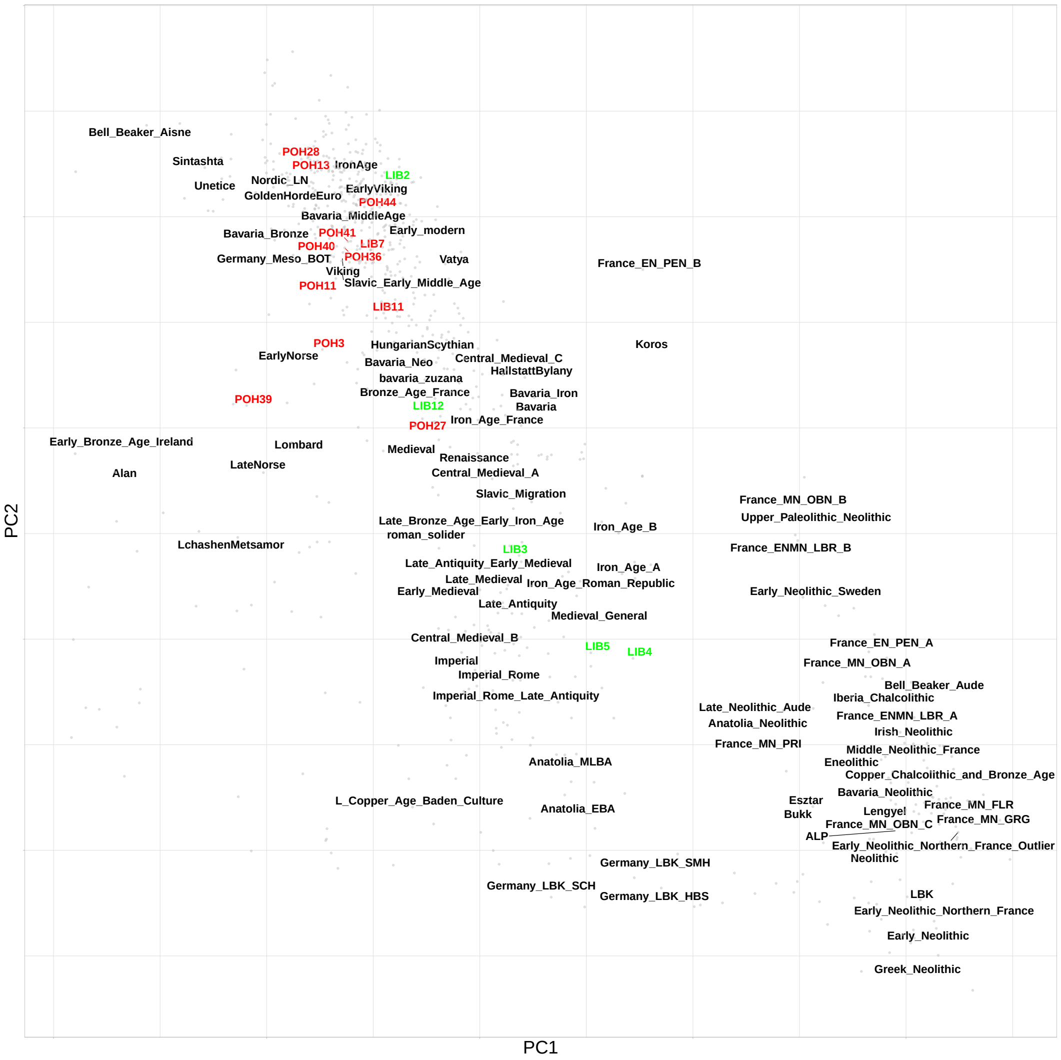

Figure 5.2: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using the plink2. Green labels correspond to Migration Era samples, red labels to Early Middle Age samples and black as reference populations. The position of each reference label is the mean PC coordinates of all individuals within that population

Northern Europe. This individual shares the most haplotypes with Viking individuals from Denmark, Estonia and the U.K. from roughly the same time period. fineSTRUCTURE analysis grouped LIB2 primarily with Viking era individuals from Sweden, Denmark, Iceland, Estonia and Norway from 300-1100AD. When painted using a set of present-day reference samples, LIB2 matches the most haplotypes and clusters with Norwegians (Fig. 5.7). Put together, these data suggests LIB2 may be a recent migrant from Viking regions.

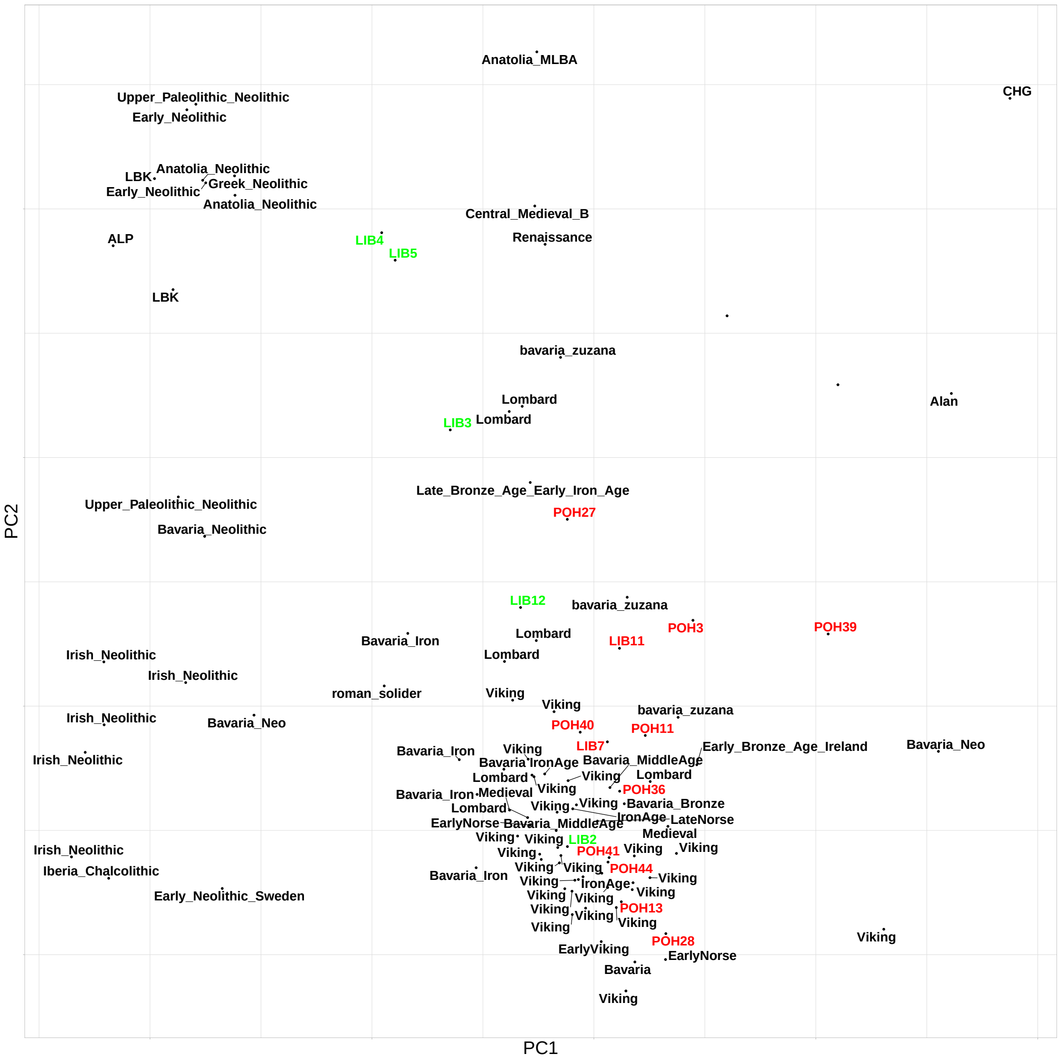

Figure 5.3: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using fineSTRUCTURE. Green labels correspond to Migration Era samples, red labels to Early Middle Age samples and black as reference populations.

There are many sources which detail the links between the Viking and Slavic peoples towards the end of the first millennium [181, 182]. However, most evidence suggests these links occurred later than the estimated radiocarbon date of LIB2. For example, it is known that the Scandinavian colonists settled in present-day Russia as early as 750. Therefore, we could suggest that this is evidence of an earlier link than previously known. In their large-scale study of ancient DNA of Viking samples from across Europe, Margaryan et al (2020) present Viking samples and ancestry in Estonia, but not until the beginning of the 8th Century, some 200 years after the estimated date of LIB2.

On the other hand LIB4 and LIB5, and to a lesser extent LIB3, show an increased affinity the Neolithic / Southern European populations relative to the other Migration Period samples, indicated by their position on the linked and unlinked PCAs. Interestingly, they share the most haplotypes with several Italian Neolithic samples, despite being separated by approximately 6000 years. Despite sharing the most haplotypes with these samples, LIB4 and LIB5 are found in fineSTRUCTURE clusters with more recent samples from Italy (Early Iron Age / Renaissance), suggesting the link to Neolithic Italy may have been transmitted by more recent populations. Both LIB4 and LIB5 share the most haplotypes with one another; this and their consistent positions on PCA and fineSTRUCTURE groupings suggest they are closely related and could be from the same local population. Similarly, LIB3 clusters with Lombard samples from Northern Italy. Historical evidence cites alliances between Slavs and Lombards in the 5th century [183]. In the 'present-day' painting, LIB3 clusters with and shares the most haplotypes with present-day Tuscans.

The appearance of Southern European-like ancestry in Central Europe, most closely related to Neolithic farmers, into the first millennium is similar to a signal found in a study exploring the ancestry of individuals with elongated skulls in medieval Bavaria (approximately 500AD) [184]. It was shown that particular individuals harbour substantial Southern-European ancestry from outside of Bavaria, closest to individuals from present-day Greece and Turkey. There are at least two possible explanations for the presence of this ancestry in the Migration Era samples. Firstly, LIB3, LIB4 and LIB5 may be similar migrants to the region. This is consistent with the fact that (at least LIB3, need to check others) is female; Veeramah et al (2018) showed that there was a tendency for females to migrate from southern regions, perhaps related to the formation of strategic alliances. Alternatively, it is possible these individuals are a leftover from a historically described Lombard migration that moved from Northern Germany through Czechia, Slovakia, Hungary and ended up in Lombardia. Accordingly, this could appear as genetic similarity to present-day populations from Northern Italy. This hypothesis is supported by the clustering of LIB3, LIB4 and LIB5 with present-day Italian samples in the 'present-day' fineSTRUCTURE analysis (Fig 5.9).

Ancestry proportion estimation using SOURCEFIND showed that the cluster containing LIB3, LIB4 and LIB5 shares 25% of their ancestry most recently with people from Anatolia, 16% from LBK (Linearbandkeramik) and 12% from a cluster containing Lombard individuals.

I performed MOSAIC admixture modelling using present-day samples as surrogates and the clusters of newly sequenced ancient samples as targets. I did not detect an admixture even when targeting LIB3, LIB4 and LIB5. This could be due to low power or a low number of samples, or that the samples are unadmixed with respect to the surrogate populations.

Finally, LIB12 displays ancestry which is more typical of the preceding Central European Bronze Age. It copies the most haplotypes from samples from Bronze Age Ireland (Rathlin) and Bavaria and is found in a cluster with several other Bronze Age samples. This suggests it may represent a 'leftover' from a local Bronze Age population which was unaffected by the Antiquity / Iron Age migrations to the region.

5.3.2 Early Middle Age Slavs represent a relatively homogeneous group typical of European Middle Ages

In comparison to the 5 Migration Period ancient Slavs, the 12 Early Middle Age Slavs (741-956 AD) represent a more homogeneous set of samples. All 12 samples were clustered into the same fineSTRUCTURE group (named Slavic Early Middle Age II), alongside Viking/Medieval samples from Ukraine, Poland and Sweden. SOURCEFIND analysis showed that the Slavic Early Middle Age II cluster derives roughly equal parts of ancestry from the clusters Viking 10C Scan I, BronzeAge I and Lombard mixed cluster. Interestingly, these are 3 ancestry sources which are similar to those identified by SOURCEFIND analysis in the Migration Period samples. I tentatively therefore suggest that the Early Middle Age Slavs were formed from the mixture of 'Northern' (represented by Viking) and 'Southern' ancestries (represented by Lombards) onto a substrate of local Bronze Age populations. Note that I suggest that these are the most representative populations and not necessarily the 'true' populations that mixed.

MOSAIC admixture modelling on the Early Middle Age samples using ancient surrogates proved inconclusive. However, using present-day individuals as surrogates provided cleaner results. The best fitting model was a 3-way admixture event involving sources closest to present-day day North-Central Slavs (76.6%), Southern-Eastern Slavs (21.9%) and East Asians, best represented by Mongolians (1.5%) (Fig. 5.4). This admixture event was estimated to have occurred 9.4 (2.5% 5.7gens - 97.5% 17.9gens) generations before the samples (Fig. **??**).

This admixture event is consistent with a signal inferred in both present-day day Eastern European individuals [15, 144]. In previous studies, this admixture event was dated to approximately 1200CE (MOSAIC) and 438CE (GLOBETROTTER). Despite the differing dates, the proportion of ancestry is consistent across studies (approximately 2%), suggesting the signal is genuine. To further support the event, the proportion of ancestry from this source is consistent across 2-way and 3-way MOSAIC admixture models.

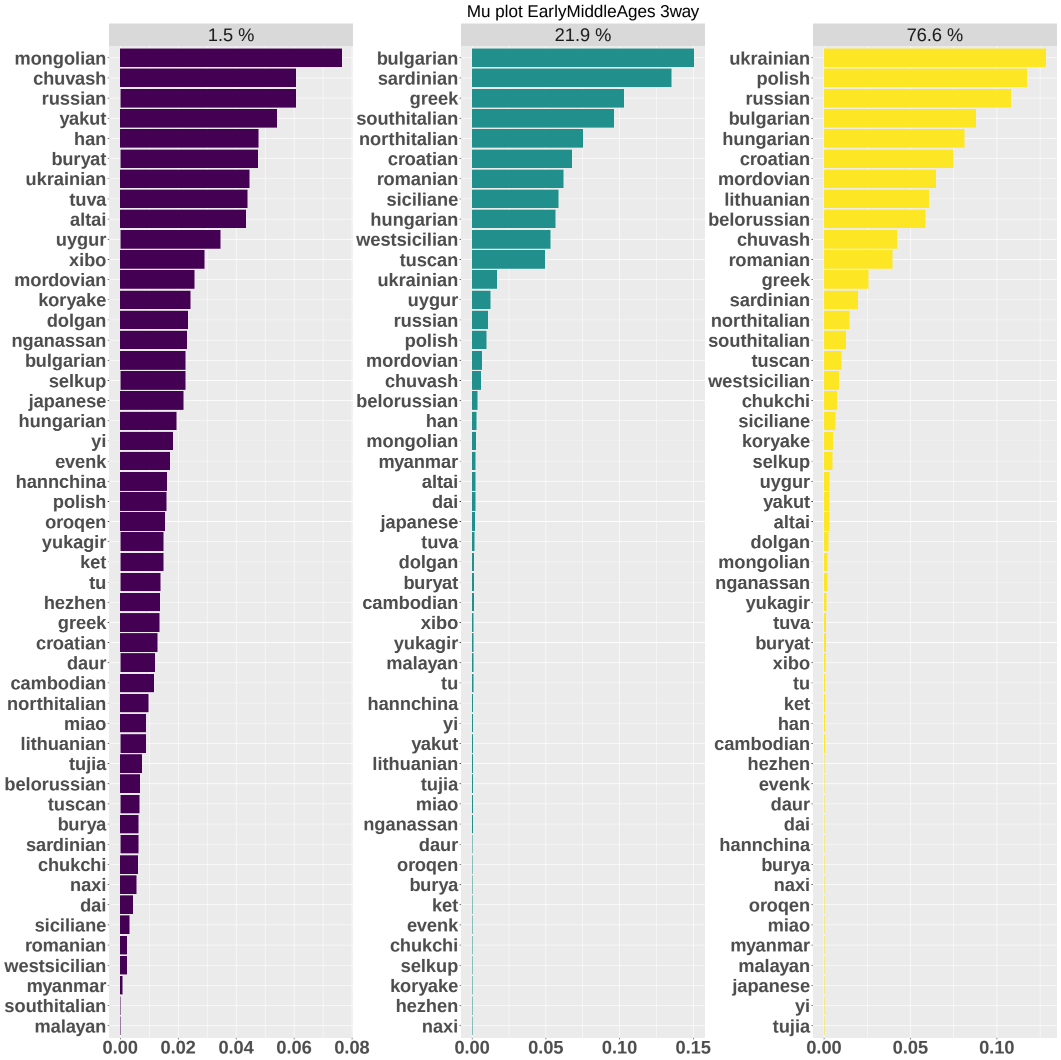

Figure 5.4: Copying matrix plot for sources in 3-way admixture event for Early Middle Age ancient Slavic samples. Each panel represents one of the 3 putative mixing sources. Labels above each panel gives the proportion that mixing source contributed to the Early Middle Age samples. Length of the bars within each panel represent the amount that putative mixing source copied from a particular population.

5.3.3 Do the samples cluster together - TVD permutation test

fineSTRUCTURE analysis suggested that the Migration Era and Early Middle Age samples did not originate from the same source population. To formally establish whether the Early Middle Age and Migration Period samples cluster within their respective population to the exclusion of the other, following Leslie et al 2015 [90], I performed a TVD permutation test. TVD is a distance metric which can be calculated from the chunklengths matrix and

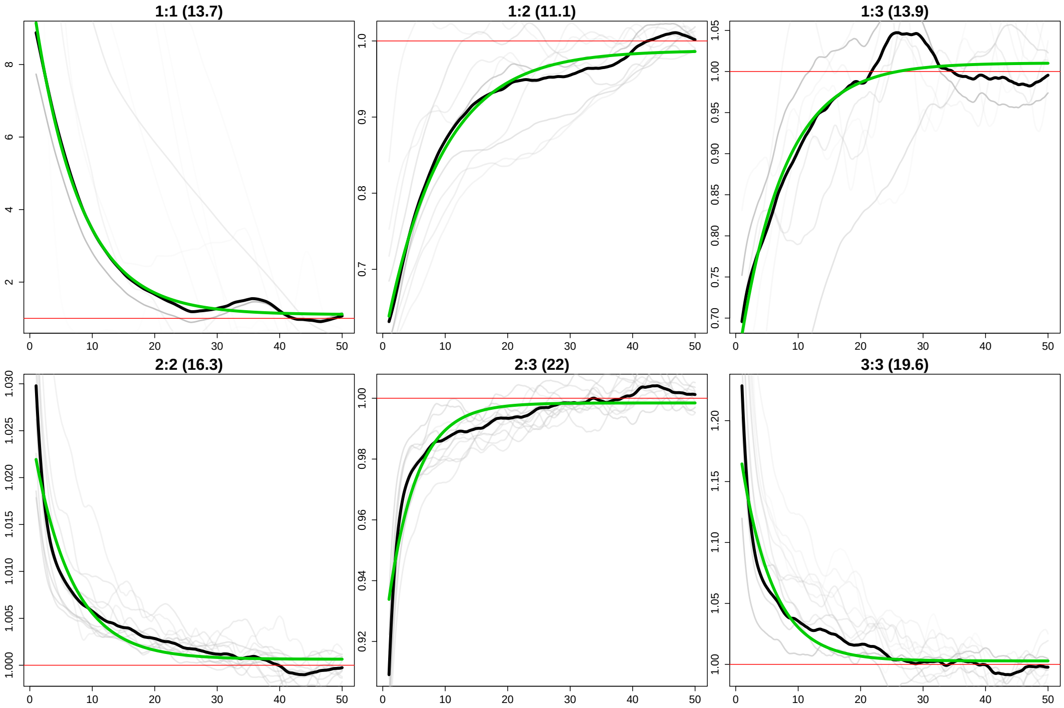

Figure 5.5: Inferred Coancestry Curves obtained from modeling Early Middle Age samples as a 3-way mixture of present-day individuals. Black lines are empirical coancestry curves across all target individuals, light grey are per individual, green is the fitted single-event coancestry curve. Note to self - need to figure out what the numbers mean but doesn't say in the manual anywhere.

is equivalent to finding the absolute distance between two copyvectors, with larger values meaning two samples have more different ancestry profiles. Full details of TVD are outlined in Appendix section B.3.

Using the ancients chunklengths matrix, I grouped the samples into Migration Period and Early Middle Age and calculated the average copyvectors C_{mp} and C_{ema} across samples within each groups. Then, I calculated the empirical TVD between the two groups as $TVD_{mp,ema} = \sum |C_{mp} - C_{ema}|$. For 10,000 iterations, I then randomly permuted the population labels among the samples and then calculated a 'random' TVD, $TVD^{rand}_{mp,ema}$ between the samples with randomly permuted populations. We can then calculate the p-value that we can reject the null model of no significant differences between the groups (not sure if this is the right way of wording it) as the number of randomly permuted iterations where $TVD^{rand}_{mp,ema} > TVD_{mp,ema}$. This test supported clustering the samples into their respective groups ($p = 0.0013$).

5.3.4 Interactions between the two groups

The previous section suggested that individuals from the Migration Period and Early Middle Ages had differing ancestry signals.

To determine the extend of mixture and continuity between the Migration Period and Early Middle Ages, I modelled each Early Middle Ages sample as a mixture of other ancients, including individuals from the preceding Migration Period. The proportion of ancestry the individuals derive from the Migration Period clusters could be used as a proxy for the degree of continuity. The proportion of ancestry derived from the Migration Period was low (mean 3.4% , range 0.4% - 12.5%), suggesting that there was a relatively large scale population replacement between the two different time periods.

5.3.5 Legacy of Slavic migrations in present-day individuals

To understand the genetic legacy the newly sequenced ancient samples left in different European populations, I painted each sample using the HellBus dataset of present-day individuals. This dataset contains a diverse set of European populations - particularly those from present-day Slavic speaking countries (Polish, Croatian, Bulgarian, Belorussian, Ukrainian, Russian) but also neighbouring non-Slavic speaking countries (Romanian, Lithuanian, Germany and Mordovia).

Principle component analysis (PCA) of the chunklengths matrix, where present-day European samples acted as donors, shows genetic similarity between ancient Slavic samples from the Early Middle Ages and present-day day Slavic speaking people (Fig. 5.6). The samples primarily cluster with present-day Polish and Belorussian individuals, but appear to fall on a cline of genetic similarity between Russians and Southern Europeans. This cline could be mediated by the possible historical admixture event between a source closest to present-day East Asians and a second closest present-day Southern Europeans, with the position of the samples along the cline dependent on the level of admixture from the different sources.

As with the ancients PCA, Migration Era Slavs are spread across the PCA. 3 samples, LIB3, LIB4, and LIB5 cluster with present-day Italians, consistent with deriving a substantial ancestry component from Southern-European sources. LIB4 and LIB5 appear to be positioned closer to Southern Italians and Greeks, whereas LIB3 is closer to Northern Italian and Tuscan populations. LIB2 shows a strong affinity to present-day Norwegians, suggesting it may be a recent migrant from Viking regions.

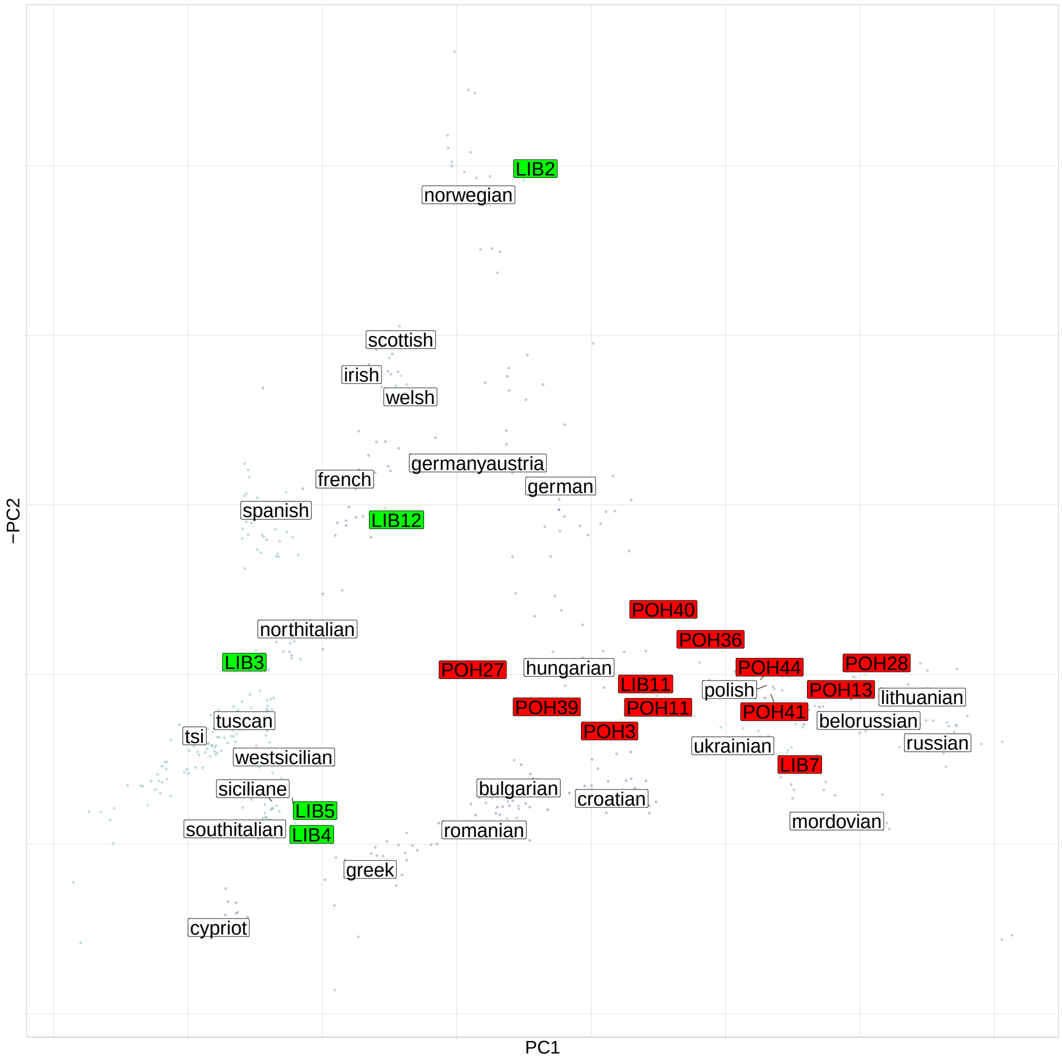

Figure 5.6: Principle component plot of newly sequenced ancient samples and reference modern individuals performed using the finestructure library. Green labels correspond to Migration Era samples, red labels correspond to Early Middle Age samples and white labels correspond to reference populations. The position of each reference label is the mean PC coordinates of all individuals within that population. Transparent coloured points correspond to present-day individuals.

The same pattern can be observed on the raw copyvector output matrix (Fig. 5.7). The Migration Era samples appear not to show any excess affinity to present-day day Slavic populations. The two samples who in previous analysis showed a strong genetic relationship to the Neolithic, LIB4 and LIB5, shared the most haplotypes with present-day day Greek individuals. This should not be surprising given present-day day Greeks have a relatively high proportion of Neolithic ancestry relative to other European populations [185].

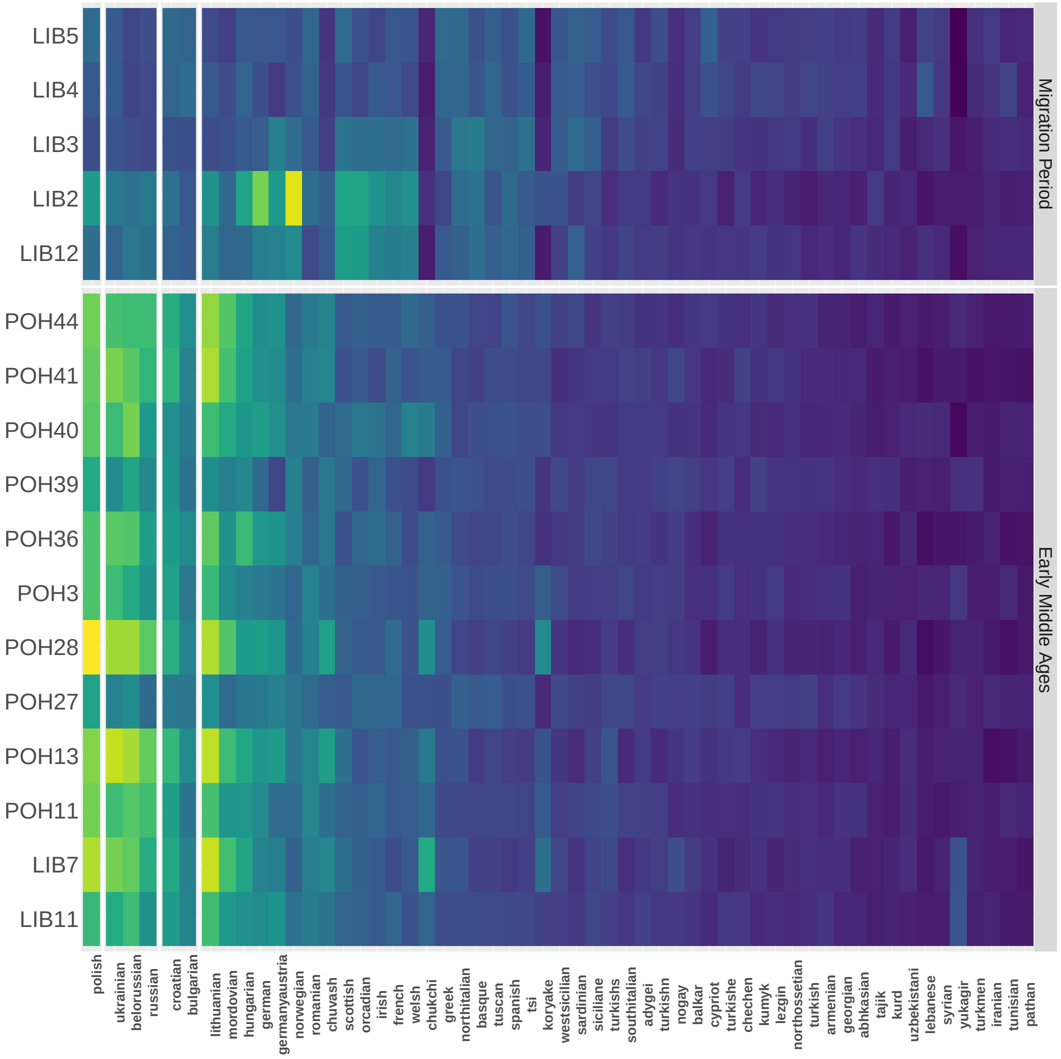

Figure 5.7: Raw chunklengths matrix from the ‘present-day’ painting. Rows correspond to different ancient recipient individuals, grouped into Migration Period and Early Middle Age period, and columns to different donor populations. Colour of cells corresponds to the total length of genome that a given donor individuals donates to that recipient, with dark/blue indicating less sharing and light/yellow colours indicating more sharing.

In contrast, the Early Middle Age samples showed a strong affinity to present-day day Slavic populations. In particular, we find that samples copy many more haplotypes from present-day day Polish individuals than they do from other populations. This is consistent with previous findings based on uniparental markers. There was also a strong affinity to several non-Slavic speaking present-day populations - notably Lithuania and Mordovian. SOURCEFIND analysis provided a qualitatively similar pattern.

To confirm that the observed results were not a result of phasing or imputing ancient individuals using present-day samples, I utilised f_3 statistics, which were performed on non-imputed genotypes. Specifically, I calculated f_3, or the branch length / amount of shared drift, between a set of present-day test populations and the grouped Early Middle Age samples. The results are qualitatively similar to those obtained using haplotype-based methods, with Early Middle Age ancient Slavic individuals being closest to samples from Eastern Europe (Fig. 5.8). However, the f_3 results do not appear to show the same degree of geographical structure; for example, Early Middle Age have a more positive f_3 with present-day Irish individuals than with present-day Croatians.

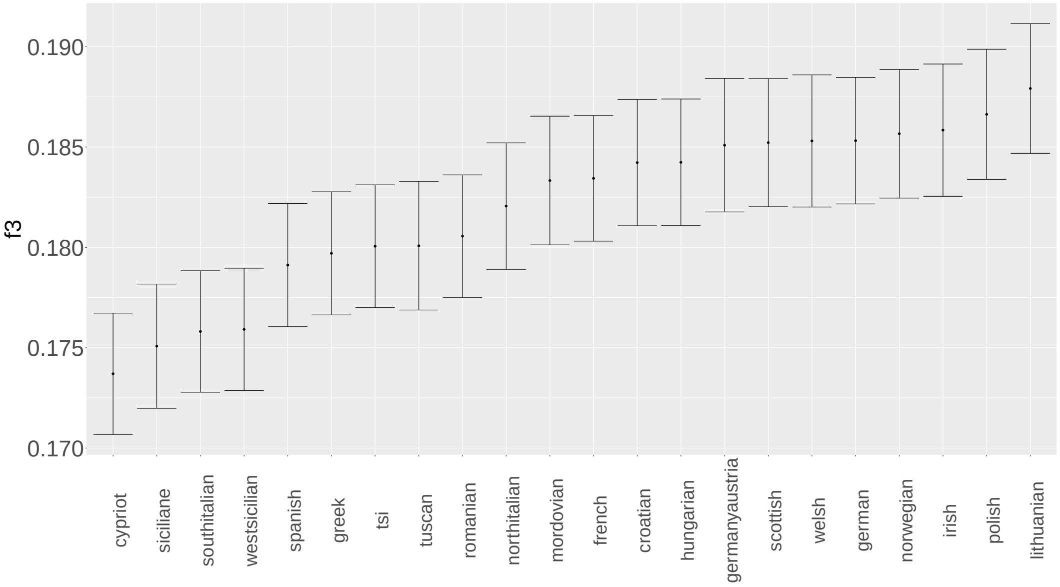

Figure 5.8: f_3 statistics in the form of $f_3(EMA, present-day; mbutipygmy)$, where *present-day* is different present-day European population. Error bars rerpesent $\pm * 2$ standard error.

5.3.6 Continuity with present-day day Slavs

The previous section strongly suggests at least some degree of continuity between Early Middle Age samples and present day Slavic populations that is not shared with the samples from the Migration Period, as the Early Middle Age samples share many more haplotypes with present-day Slavs compared to Migration Period.

To explicitly test the hypothesis that the Early Middle Age samples were continuous with the present-day day Slavic populations, I used *qpWave*, which tests the number of streams of ancestry from a set of *right* populations into a set of *left* populations, $qpWave(left =$

$croatian, lithuanian, polish, ukrainian, right = middleage, migration$). The matrix with rank $r = 0$ can be rejected ($p = 0.112$). Note - not sure how to interpret this.

5.3.7 Genetic structure and admixture events of present-day Slavic people

As described in the introduction, several studies have investigated the structure of present-day Slavic populations, but none have integrated autosomal DNA from present-day and ancient samples and analysed them jointly with haplotype-based methods. I performed an all-v-all painting of a selection of present-day European populations and all newly sequenced ancient Slavic samples and applied the fineSTRUCTURE algorithm to the resulting chunkcounts matrix. fineSTRUCTURE generated 32 clusters.

Present-day Slavs do not form a monophyletic group within the fineSTRUCTURE dendrogram to the exclusion of non-Slavic populations (Fig. 5.9), as several non-Slavic speaking populations such as German, Irish and Scottish cluster in the main clade containing Slavic speakers. Within present-day Slavs, structure is apparent; speakers of ‘Southern’ Slavic languages from Croatia and Bulgaria form a group to the exclusion of ‘Eastern’ Slavic speaking populations from Belarus, Russia and Ukraine. Individuals from Poland cluster with ‘Eastern’ Slavic speakers, suggesting the principle axis of variation splits populations into ‘North-West’ and ‘South-East’ groups.

Of the Early Middle Age samples, 3 samples (POH3, POH39, POH27) were present in the ‘South-East’ Slavic cluster, falling into a group composed of Bulgarian and Romanian samples. The remaining 7 samples are found in the ‘North-West’ cluster containing samples from Lithuania, Poland, Ukraine and Belarus. Painting the samples using present-day individuals has thus uncovered structure that was not able to be detected by looking only at ancient samples. It also suggests the structure of Slavic populations into was present at least as early as the date of these samples.

Previous studies have identified admixture events in present-day Slavic populations involving an East Asian source approximately 440 to 1080 CE [144, 186]. In previous sections, I showed that this signal exists in the Early Middle Age ancient samples and is best characterised by populations from present-day Mongolia (Fig. 5.4).

I employed MOSAIC [144] to replicate these results and determine whether a similar admixing source is present in the ancient populations.

When considering 2-way admixture event, all of the tested populations, bar the Migration

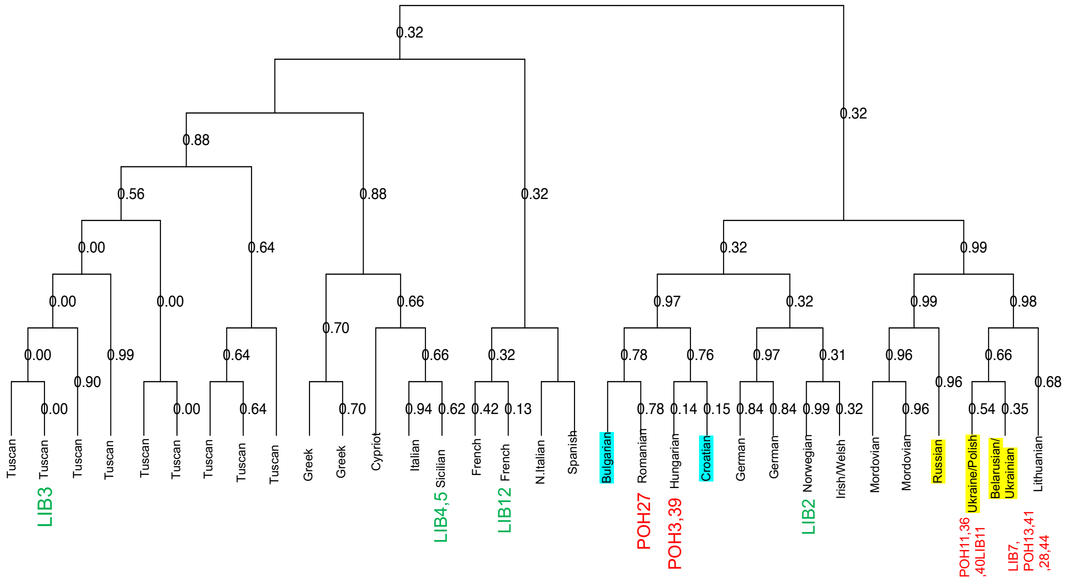

Figure 5.9: Population dendrogram generated by the fineSTRUCTURE tree building algorithm. Labeled tips refer to the primary population(s) represented in that clade. present-day non-Slavic populations shown in black. 'South-East' Slavs highlighted in cyan and 'North-West' Slavs highlighted in yellow. Migration period individuals superimposed in green and Early Middle Age samples superimposed in red. Read fineSTRUCTURE paper for description of edge values. Note: some tips contained more than one population but were not included as labels to save space.

Period Slavs, showed evidence of an admixture event involving a minor source which has the lowest Fst with present-day Uygurs. The dates and bootstrapped confidence intervals are given in Fig. 5.10. Other than Norwegians and Croatians, whose estimated dates are later and earlier respectively, the admixture dates for other populations appear to be constrained to approximately 1250 CE. This date is similar, but slightly later than that obtained from Hellenthal et al (2014), who estimate it to be 440 to 1080 CE.

Interestingly, most present-day Slavic speaking populations, such as present-day Polish, show evidence of a 3-way admixture event, where the middle component has the lowest F_{st} with Migration Era ancient samples (Fig. 5.11). The major component has a low F_{st} with Early Middle Age Slavs. This suggests that the formation of present-day Slavic populations could have occurred via an admixture event(s) involving Migration Era individuals with high levels of Southern European ancestry, Middle Age Era samples which show a strong affinity to present day Eastern Europeans, and a small but significant East Asian source best represented by present-day Uygurs.

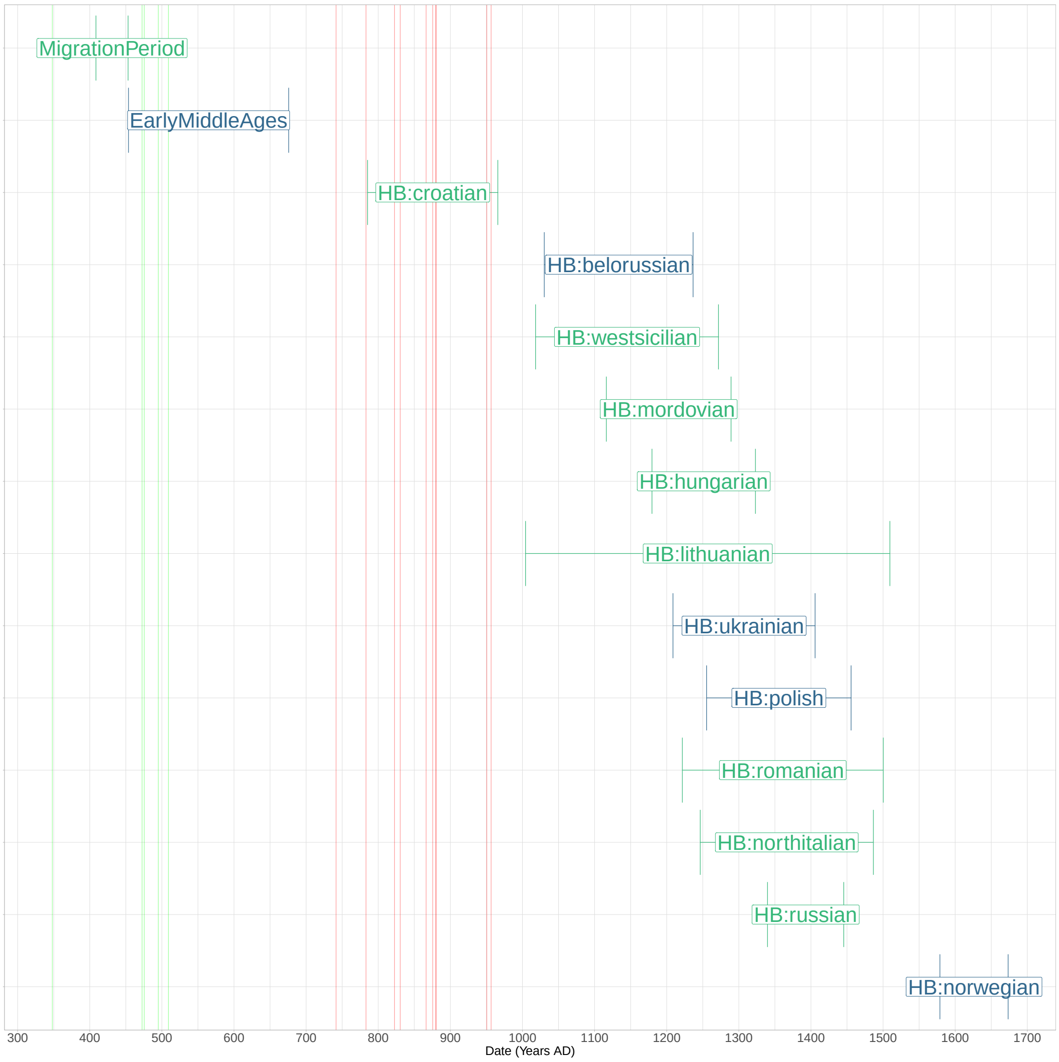

Figure 5.10: MOSAIC inferred 2-way admixture dates with bootstrapped 97.5% and 2.5% CI. Vertical green lines correspond to radiocarbon estimated dates of Migration Period samples and red lines equivalent for Early Middle Age samples. Estimated dates obtained by assuming an average generation time of 26 and date of birth of 1950 for present-day samples.

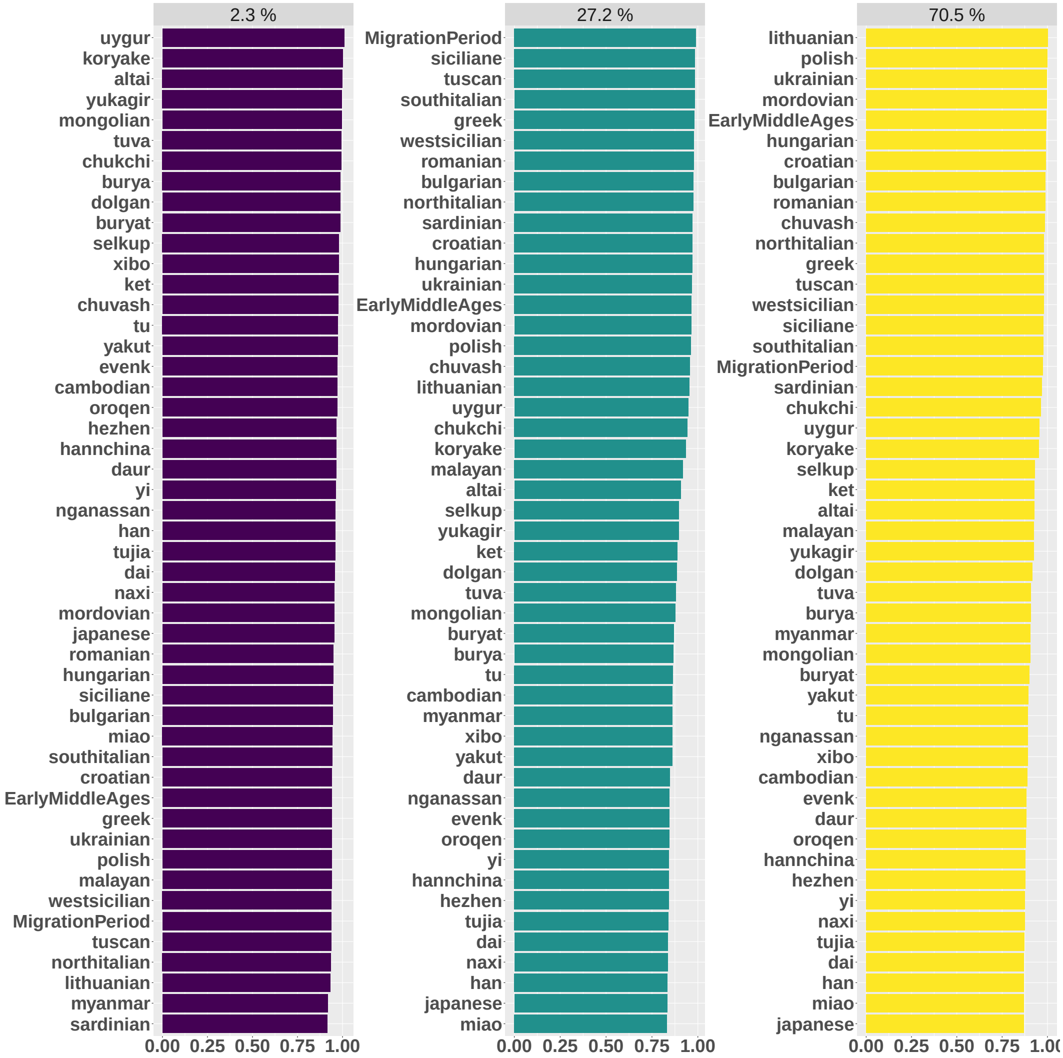

Figure 5.11: $1 - F_{st}$ between 3 inferred mixing sources for present-day Belorussians. Each panel represent a different mixing source. Each bar gives the value $1 - F_{st}$ between that samples population and the mixing source. Higher values of $1 - F_{st}$ suggest that source is well represented by a particular population.

5.4 Discussion

The combined results from the Migration Period suggest the individuals living in Czechia during this time period were of mixed ancestry and did not originate from the same source population. The diverse set of ancestries, spanning from Scandinavia to Southern Europe imply that the Migration Period was truly a period of Migration where individuals from distal ends of Europe lived among one another. In particular I inferred ancestry sources from Southern Europe and Scandinavia.

The results from the analysis of combined ancient and present-day genomes are consistent with those from Kushniarevich et al (2015) [172] who determined that Eastern (Russia, Belarus, Ukraine) and Western (Polish) central European Slavs form a cluster to the exclusion of Southern Slavs (Croatia, Bulgaria), whilst also remaining distinct from geographically proximate Germanic (German/Austrian) and Baltic (Lithuanian) populations. This is also consistent with results from Veeramah et al 2011, who showed that Sorbs, a west-Slavic population found between Poland and Germany, have a much stronger affinity to more distant Slavic populations from Czechia than to more proximate Germans [138]. Similarly, I inferred that the Slavisiation of the Balkan peninsula doesn't extend beyond Croatia; the cluster of Croatian individuals only derives 1.2% of their ancestry from nearby Greek sources. However, admixture modelling suggested that Southern Slavs show signals of a historic admixture event where the minor source is related to present-day Mediterranean populations. An admixture event with a similar minor source is inferred in Migration period samples, albeit dated further in the past.

I recapitulated a previously described admixture event into not only present day Slavic speaking populations, but also Southern Europeans (e.g. North Italians). The source of this East-Asian admixture is closest to present-day Uygurs. However, the true ancient population that was responsible to transmitting East-Asian ancestry into Europe is yet to be determined. It seems likely that the ancestry was brought to Europe via an intermediate population containing East Asian ancestry, such as the Huns or Turkic peoples.