Chapter 4

Bavaria ancient DNA

4.1 Introduction

Throughout the Pleistocene and Holocene, Germany has been the setting for many population movements and admixture events of modern humans. The Swabian Alps is home to one of the earliest symbolic art, dated to at least 32kya [121] and musical instruments dated to 40kya [122], both assigned to the Aurignacian tradition.

Later, the region was also home to one of the first Neolithic traditions in the *Linearbandkeramik* (LBK), a key culture in the Neolithisation of Europe. Early LBK populations across Germany mixed with the preceding Mesolithic hunter gatherer populations [79, 123–126]. At the end of the Neolithic, a new ancestry was detected [79, 127] in concert with the arrival of the Corded Ware Complex [128]. The new ancestry was most closely related to the Yamnaya Pastoralists from the Pontic Capsian Steppe. Recent studies using ancient DNA samples have shown that the arrival of Steppe-related ancestry occurred no earlier than 2700BC [129] and followed a steep rise shortly after.

The beginning of the Bronze Age also saw the arrival of the Bell Beaker culture across Europe.

Cherry-Tree cave (Kirschbaumhöhle in German) represents an unique opportunity to study the transect of samples from the Neolithic to the present-day. The cave represents a relatively untouched layer of stratigraphy.

In the present-day, Germany represents a boundary point between East and West Europe, with a relatively sharp genetic boundary occurring between Germany and Poland to the east [130–132]. However, within Germany, studies have shown that there is only very weak

substructure [133]. Questions remain as to the origin of this East-West structure; is it recent structure, or does it persist to the Middle Ages or earlier?

Here, I present novel data from 11 medium-to-high coverage samples from two sites from Southern Germany and one site from one from Southern Austria. In particular, the samples from Kirschbaumhöhle span from the Late Neolithic to the Iron Age, providing an opportunity to study a time transect in a narrow geographic region.

Previous studies into the genomic history of Bavaria have focused, for example, on the mixed ancestry of migrant females during the Early Middle Ages.

A collaborator, Prof. Joachim Burger, Johannes Gutenberg University Mainz, posed the following 3 questions.

1. **Second Neolithic immigration wave**. One of the samples is thought to have belonged to the first wave of farmers carrying farming technology from the near-east to Europe, and another to the second. Do we observe genetic differences between the two waves of samples?

2. **Cherry Tree Cave**. How can we make sense of the genetic ancestry changes from the Late Neolithic through to the Iron Age in Cherry Tree Cave? Do we see evidence of genetic continuity between the ages and are they characterised by admixture from outside sources?

3. **Germanic / Slavic divide**. Is there a distinction between the Germanic and Slavic Middle Age samples? How do these populations compare to the preceding samples from the Bronze and Iron ages?

4.2 Methods

4.2.1 Data generation

Eleven whole-genomes of ancient individuals were generated by collaborators at the Johannes Gutenberg, University of Mainz, Germany. The estimated radiocarbon dates range from 1060AD to 5200BC (Fig. 4.2). Six of the samples were found in Cherry-Tree Cave in the Bavarian district of Forchheim (Fig. 4.1), four from futher South in the region of Dingolfing/Essenbach and one sample from southern Austria. The samples had a median coverage of 4.84x and ranged from 0.7x to 17.52x. Full details of coverage, location and dates are given in Table 4.1.

I was given the data of each newly sequenced sample in `vcf` format.

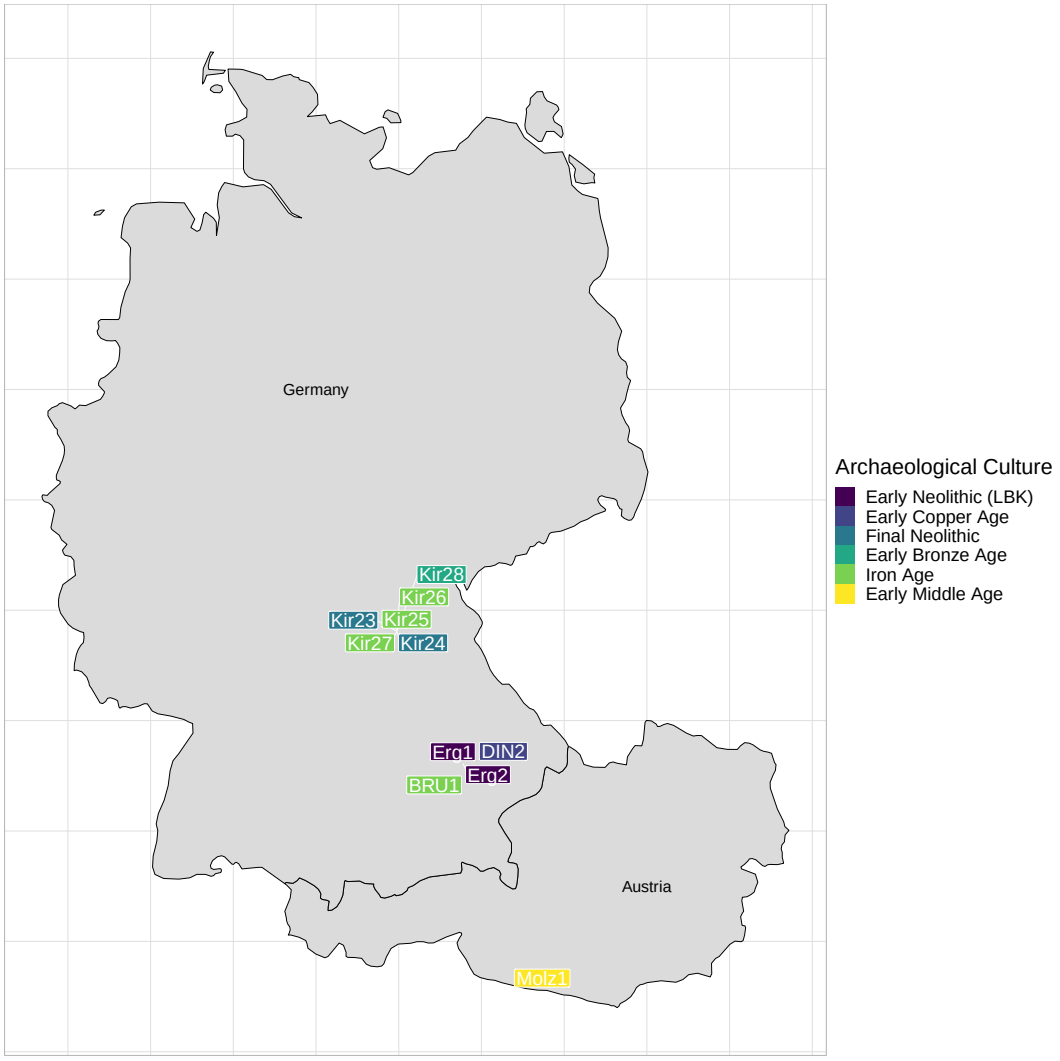

Figure 4.1: Map of newly sequenced ancient individuals, positioned according to where they were excavated. Colour on label corresponds to archaeological culture which they were found.

4.2.2 Genotype imputation and phasing using GLIMPSE

In order to compare the genetic variation in the newly sequenced samples to a reference dataset, I merged them with the 942 ancient samples from the literature detailed in Appendix section A.1, resulting in a total of 955 samples in `.bcf` format with genotype likelihood data at 77,213,942 genome-wide SNPs. Data was then split into separate `.bcf` files for each chromosome and indexed using bcftools [134].

As ChromoPainter analysis requires each ancient sample to be phased, I followed the recommended GLIMPSE [64] imputation and phasing pipeline (`https://odelaneau.github.`

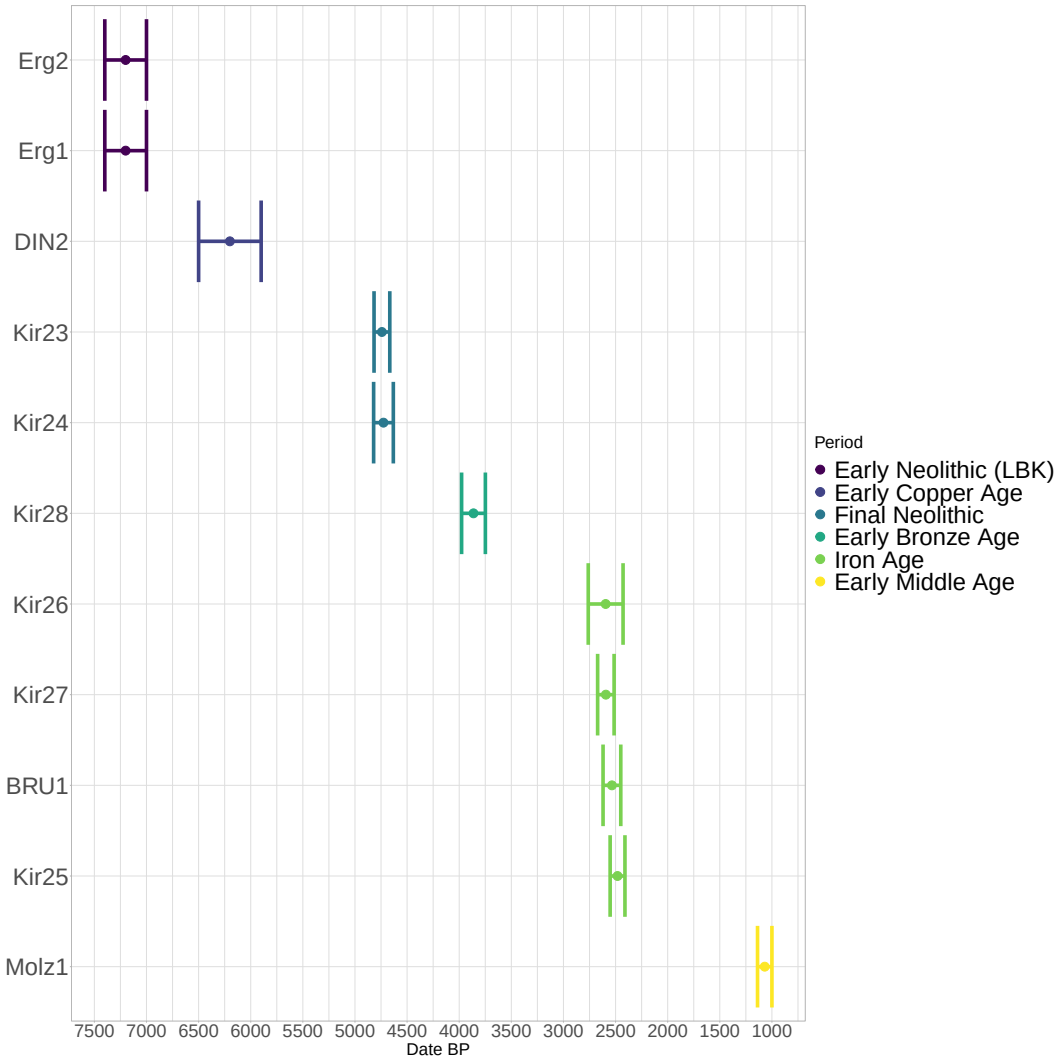

Figure 4.2: Estimated radiocarbon dates for each newly sequenced ancient individual, grouped by archaeological period.

Sample.ID	Location	Date	Period	Coverage
Erg1	Ergoldsbach-Essenbach	5200	Early Neolithic (LBK)	4.52
Erg2	Ergoldsbach-Essenbach	5200	Early Neolithic (LBK)	0.71
DIN2	Dingolfing	4200	Early Copper Age	1.71
Kir24	Cherry Tree Cave	2762	Final Neolithic	3.98
Kir23	Cherry Tree Cave	2741	Final Neolithic	17.52
Kir28	Cherry Tree Cave	1863	Early Bronze Age	17.30
Kir26	Cherry Tree Cave	595	Iron Age	4.84
Kir27	Cherry Tree Cave	593	Iron Age	16.60
BRU1	Bruckberg	535	Iron Age	11.54
Kir25	Cherry Tree Cave	481	Iron Age	4.55
Molz1	Molzbichl	1069	Early Middle Age	13.22

Table 4.1: Table providing details for the newly sequenced Bavarian samples.

io/GLIMPSE/tutorial_b38.html), using the 30x-coverage 1000 genomes dataset [76] as a reference panel. This resulted in phased haplotypes and posterior genotype likelihoods for each of the 955 individuals.

4.2.3 Determination of uniparental haplogroups

To determine the mtDNA and y-chromosome haplogroups for each newly sequenced ancient sample, I used Haplogrep (https://haplogrep.i-med.ac.at/) [135] on the raw .fastq file for each sample.

4.2.4 IBD sharing

I used hap-IBD [136] to estimate IBD segments greater than 2cM in length between all pairs of ancient individuals above 1.5x coverage (n=466), using the phased output from GLIMPSE as input haplotypes, the genetic maps from (http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/plink.GRCh37.map.zip) and leaving all parameters as default. I estimated IBD segments for each chromosome separately and summed the lengths between each pair of individuals across all chromosomes.

4.2.5 plink PCA

To obtain a broad overview of the ancestry of the newly sequenced individuals in the context of the 942 literature samples detailed in Appendix section A.1, I performed PCA on the pre-imputation genotypes using plink2 [137]. Performing a PCA in plink2 allows for both an understanding of genome-wide variation patterns and the identification of any data quality issues that are independent of imputaiton, phasing or ChromoPainter analysis.

I retained the 500,000 markers with the lowest amount of missingness across all samples and 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.

4.2.6 ChromoPainter and fineSTRUCTURE analysis

To characterise of the ancestry of the newly sequenced ancient samples in the context of other ancient individuals, I first selected all newly sequenced samples and literature samples above 1.5x coverage (n=466) and performed an 'all-v-all' painting where each sample was painted using all other samples. 1.5x was somewhat arbitrarily chosen as my previous work has shown this is a suitable threshold for the inclusion of samples for ChromoPainter analysis (section 2.6.4); whilst I show 0.5x as the cut-off for coverage-related effects, I chose to be

Population	Number of Samples
HB:tsi	196
HB:spanish	68
HB:bulgarian	62
HB:german	60
HB:french	56
HB:russian	50
HB:greek	40
HB:ukrainian	40
HB:croatian	38
HB:hungarian	38
HB:norwegian	36
HB:southitalian	36
HB:polish	34
HB:romanian	32
HB:mordovian	30
HB:cypriot	24
HB:northitalian	24
HB:lithuanian	20
HB:siciliane	20
HB:westsicilian	20
HB:belorussian	18
HB:tuscan	16
HB:irish	14
HB:scottish	12
HB:germanyaustria	8
HB:welsh	8

Table 4.2: Name of population and number of samples used in the present-day ChromoPainter analysis

conservative and opt for a higher threshold. I used this painting, hereafter referred to as 'ancient' painting, to perform fineSTRUCTURE clustering and tree building on ancient samples.

I performed Principle Component Analysis on the coancestry matrix of the 'ancients' painting using the `prcomp_irlba` function from the irlba R library. To account for the fact that the diagonals of the co-ancestry matrix are always zeros (as an individual cannot be painted by themselves), I set the diagonal of each row to be the mean of that row. Although there were 466 individuals in the 'ancients' painting, not all of these were included in the chunklengths PCA. This was because many individuals in that set were not relevant to exploring the ancestry of the Bavarian individuals. For instance, when plotted, samples such as those from the Xiong Nu, a 3rd century BC culture from inner Mongolia, dominate the variation in a PCA to the point where identifying structure between the samples of interest becomes challenging. Therefore I removed 327 individuals based on visual inspection of the first two principal components

To determine the genetic similarity between the newly sequenced ancient samples and

present-day populations, I performed an 'all-v-all' painting using a selected group of 26 present-day European populations (Table 4.2) from the HellBus dataset (described in appendix section A.4) and the 11 newly sequenced ancient individuals, hereafter referred to as 'present-day painting'.

I applied fineSTRUCTURE (v0.0.5) [14] to cluster the chunkcounts ChromoPainter output for the 'ancients' painting. fineSTRUCTURE assigns individuals to genetically homogeneous clusters, estimates the 'true' number of clusters and builds a dendrogram of genetic similarity based on a tree-building algorithm. This is particularly useful when combining many samples from different studies, as is the case with the 'ancients' 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 group labels. fineSTRUCTURE 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.

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`).

4.2.7 SOURCEFIND

I used SOURCEFIND [16] to infer the proportions of ancestry by which each newly sequenced ancient individual is most related to a set of surrogate populations. While this method does not explicitly attempt to identify admixture, in contrast to (e.g.) ALDER [138] or GLOBETROTTER [15], it can reflect admixture proportions [16] but more generally reflects recent ancestry sharing patterns.

The first analysis used the ancients painting and only three surrogates: Western Hunter-Gatherers, Neolithic farmers from Anatolia and Yamnaya. This analyses reflects previous research suggesting most ancient Europeans, with the exception of some Palaeolithic Hunter-Gather populations [90], descend from the mixture of three sources well-represented by these groups. The second analysis attempted to characterise more fine-scale ancestry patterns, by modelling each target ancient individual (using the same ancients painting) as a mixture of all sampled ancient populations above 1.5x coverage (n=466) that had an average sample age no more than 100 years younger than that of the target individual. The third analysis used the "modern" painting and formed each ancient individual as a mixture of all present-day populations shown in Table 4.2. For each of these analyses, I found the mean and 95%

credible interval of ancestry estimates across 2,000,000 posterior samples combined from three independent SOURCEFIND runs that each sampled every 10,000 MCMC iterations after discarding the first 10,000 MCMC iterations as "burn-in".

4.2.8 MOSAIC admixture analysis

I inferred admixture events, dates and proportions in newly sequenced ancient samples using MOSAIC, a haplotype-based method [139]. While MOSAIC cannot infer multiple pulses of admixture from the same admixing sources as GLOBETROTTER [15] can, in theory it is unlikely we would have adequate power to identify such multiple pulses when analysing only a single ancient sample, as is the case in this study. Furthermore, the 'painting' step and admixture inference step in MOSAIC are combined, providing a simpler pipeline and more flexible assignment of different surrogates relative to GLOBETROTTER (i.e. the set of surrogates can be changed without repainting the samples).

Thus, I used MOSAIC to perform two different admixture analyses. First, I performed an 'ancient surrogates' analysis where the all ancient samples above 1.5x coverage (n=466) were used as surrogates to admixing sources. I used the fineSTRUCTURE groupings to categorise ancient samples into surrogate populations.

I also performed a 'present-day surrogates' analysis where a selected set of present-day populations (Table 4.2) were used as surrogates. While using present-day populations to reflect ancestry patterns in ancient individuals may be counter-intuitive, the larger sample sizes and larger variety of present-day populations can provide more clean results relative to using ancients

I ran MOSAIC using default settings, assuming two or three admixing sources per target individual/population. For populations with more than one sampled individual, MOSAIC provided bootstrap-based 95% confidence quantiles around date estimates. MOSAIC also estimates f_{st} between the set of surrogates and the estimated 'true' mixing source, which is useful when a close proxy for the 'true' mixing source is not available

4.2.9 F-statistics

Many of the relevant samples in the literature were of very low coverage (< 0.1). As my work in section 2.6.4 indicated that samples with less than 0.5x coverage cannot reliably be analysed using ChromoPainter, I also used F-statistics [29] that are mostly robust to coverage related effects [31]. In particular I used Admixtools (`https://uqrmaie1.github.io/admixtools`) to analyse 942 individuals from 143 populations (Appendix section A.1 including many

low-coverage samples from LBK cultures presented in Rivollat et al (2020) that would not have been suitable for use with ChromoPainter [140]. This analysis also incorporated 2280 present-day individuals from 144 populations from the HellBus dataset as putative ancestry surrogates for tested ancient individuals. Populations shown in Table 4.2.

For the input to ADMIXTOOLS, I used the genotyped imputed from GLIMPSE, as it has been shown that using imputed markers reduced reference bias relative to using pseudo-haploid markers [35]. I then used the f_4 branch test to test whether two populations form a clade relative to two other populations. For example, the expected value of $f_4(french, german; yoruba, mbuti)$, which tests whether {french,german} form a clade relative to {yoruba,mbuti}, should not give a score significantly different to zero. In contrast, exchanging $french$ with $yoruba$ would yield a significantly positive f_4 scores, with strength of evidence to reject the null ($f_4 = 0$) measured using standardised Z-statistics.

I also used the f_3 test, denoted $f_3(A, B; C)$, to (i) estimate the branch length between A and B after their divergence from C, or (ii) test whether C descends from an admixture event between sources represented by A and B. The latter can occur if C has a substantial number of SNPs with allele-frequencies which are intermediate between A and B.

Finally, I used qpAdm to infer ancestry proportions, following the protocol described in Olalde et al (2018) by choosing the following populations/samples as outgroups: *Mota*, *Kostenki14*, *papuan*, *han*, *hannchina*, *mbutipygmy*, *sannamibia*, *yakut*. These outgroups were suitable for use in investigating ancient Eurasians, since they are asymmetrically related to many ancient populations, but do not show evidence of recent gene flow with them.

4.3 Results

4.3.1 Broad-scale ancestry changes in Bavaria reflect those found elsewhere in Europe

The newly sequenced samples from the Early Neolithic (approx 5200BC) and Copper Age (approx 4200BC) cluster with other samples from the European Neolithic (Fig. 4.3). Previous studies have explained the pattern observed when Neolithic samples are plotted on a Principle Component Analysis performed using plink2 [123]; the earliest Neolithic samples, from Anatolia and Greece, and who are thought to be the source population from which all subsequent Neolithic farmers derive [38, 126, 141–143], are usually positioned at the end of the cluster which is farthest away from the hunter-gatherer samples (for example, WHG on Fig. 4.3). This likely reflects the fact that they are unadmixed with respect to the later

Neolithic samples. As the Neolithic progressed, farmers from the near-east mixed with local hunter-gatherer groups in central Europe [123] and acquired local hunter-gatherer ancestry. Accordingly, these samples are shifted away from the earlier Neolithic samples towards the hunter-gatherers. With this in mind, the position of Erg1, shifted north away from the contemporaneous sample Erg2, is suggestive of hunter-gatherer admixture.

There are four key observations from the Figure 4.3 PCA regarding the new samples:

1. The two Late Neolithic individuals are genetically separate, with Kir24 positioned close to Yamnaya and Kir23 clustering with Neolithic Europeans.

2. The Bronze Age sample Kir28 clusters with other European Bronze Age samples

3. The four Iron Age samples (Kir25, Kir26, Kir27 and BRU1) cluster towards the Neolithic individuals and other European Iron Age samples

4. The three Medieval period samples (Alh1, Alh10, Molz1) cluster with the Bronze Age sample Kir28 instead of the Iron Age samples.

4.3.2 Early Neolithic

The three Early/Middle Neolithic samples all display a strong affinity to Anatolian farmers, consistent with the prevailing theory that near-eastern farmers were responsible for the spread of early agricultural technology across Europe, and that all Neolithic farmers share recent common ancestry [38, 141–143]. fineSTRUCTURE clustering grouped Erg1 with 2 samples from Upper Palaeolithic/Neolithic Italy and DIN2 with Early/Middle Neolithic samples from Germany, Greece, Anatolia and Hungary. Despite their age, the genetic variation of the Early Neolithic samples falls well within the variation of present-day individuals; when painted using present-day samples, the 3 Early Neolithic individuals cluster with present-day Italians, consistent with findings from previous research [38, 79] (Fig. 4.4). Erg1 was assigned to mtDNA haplogroup K which has been found in Neolithic and pre-pottery sites across Europe [126, 144] and Western Asia [145, 146].

Erg1 is from the *Linearbandkeramik* (LBK) culture and is speculated to have belonged to the first wave of immigrants carrying farming technology from south-eastern Europe or Anatolia into central Europe. DIN2 is from a nearby site, around 500 years more recent, and is thought to potentially belong to a second wave of farmers who migrated along the Danube. It is unclear to what extent these different waves corresponded to populations with different ancestries.

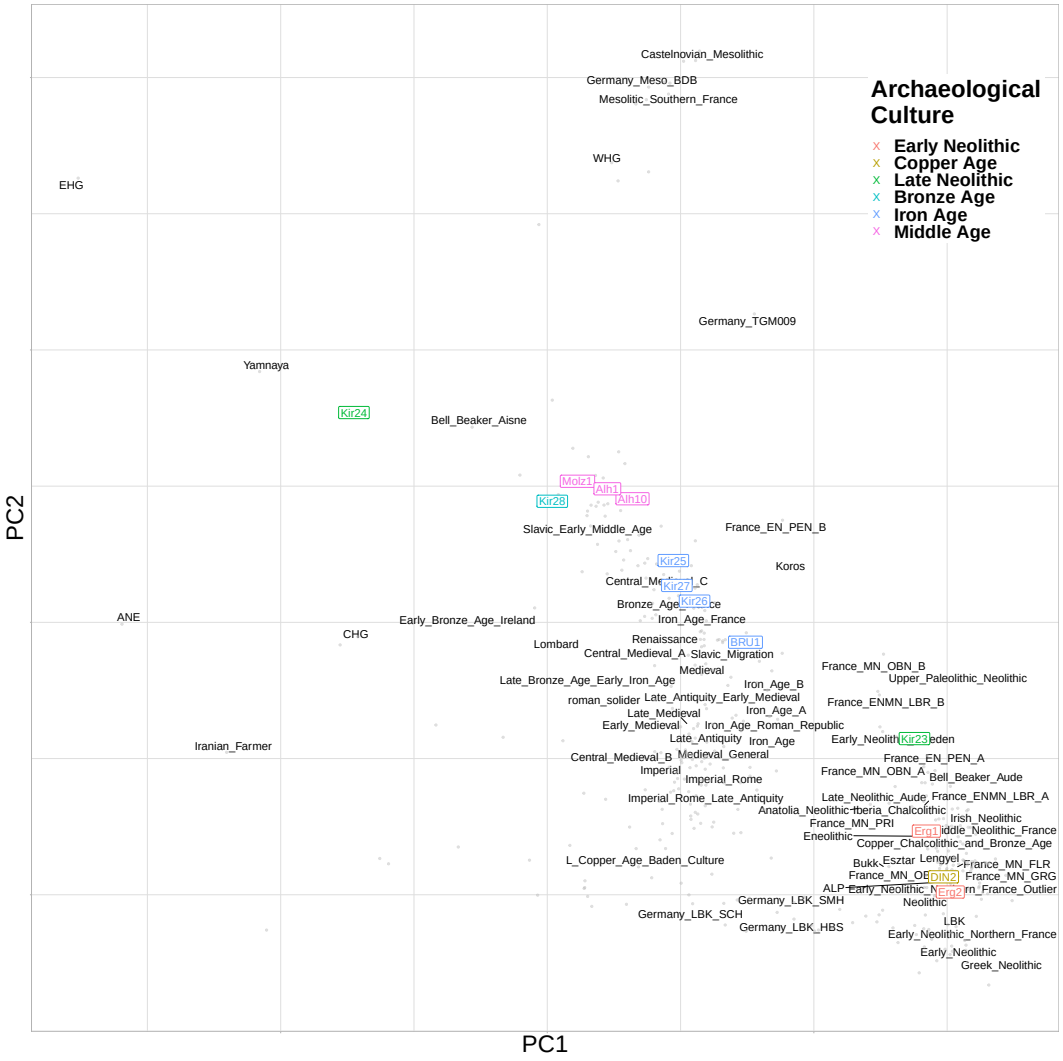

Figure 4.3: Principle component analysis of genotype matrix using plink2. Grey points indicate principle component coordinates for each sample. Black text indicated mean principle component coordinates for all individuals within that group. Coloured labels represent newly sequenced ancient samples.

When painted using 465 ancient samples from the literature and the newly sequenced samples, Erg1 had the lowest TVD (TVD is a distance metric based on ChromoPainter copyvectors; calculation and justification outlined in Appendix section B.3) with DIN2, supporting the hypothesis that they were from the same source population. Erg1 had the second lowest TVD with Ess7, another LBK sample from Essenbach, Germany. DIN2 also shares low TVD with Ess7, but has the lowest TVD with NE5, NE4 and NE7, samples assigned to Middle and Late Neolithic cultures on the Hungarian plane. DIN2 was assigned to mitochondrial haplogroup J1C, the same as the samples NE4 and NE5. Both the autosomal and mtDNA link to Neolithic Hungary supports the hypothesis that DIN2 migrated along the Danbian route.

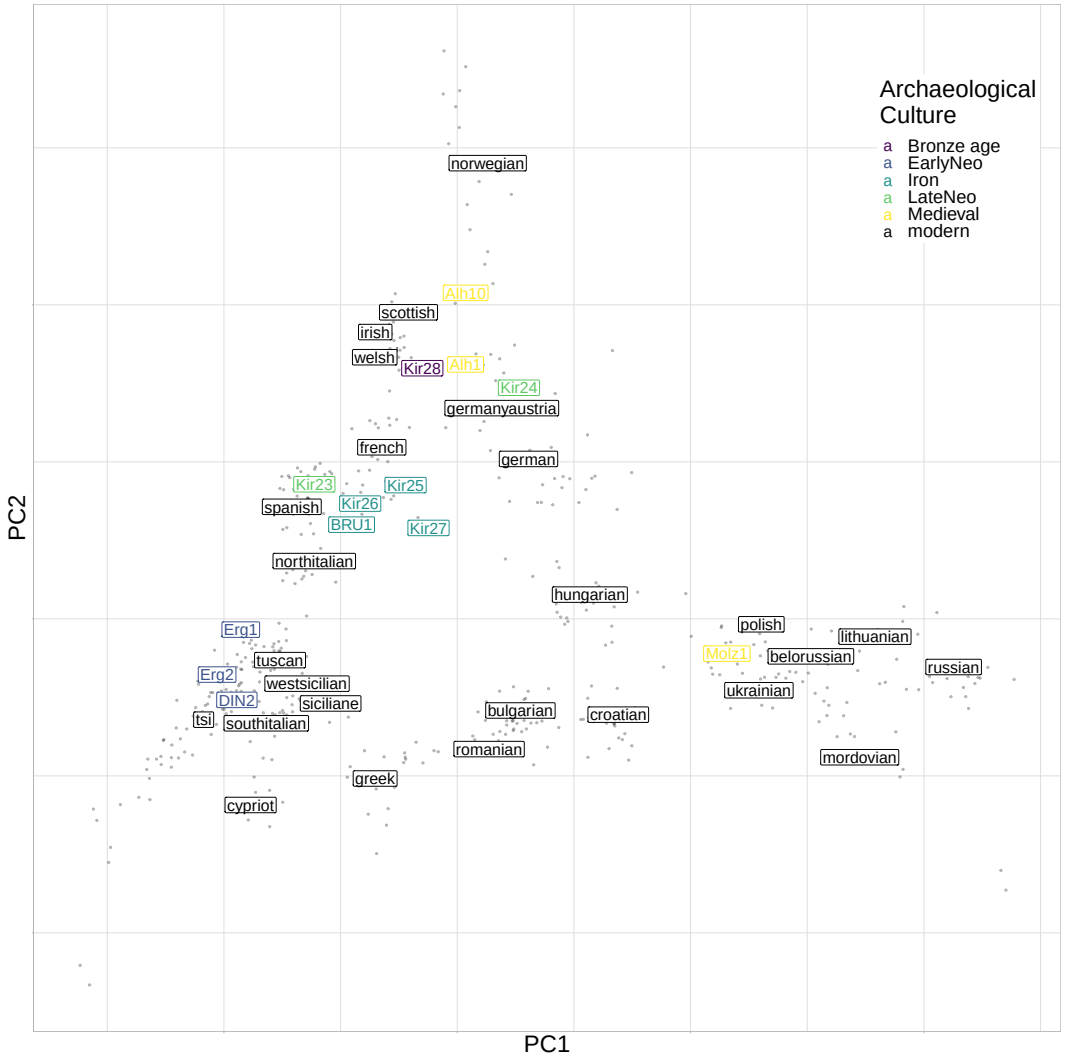

Figure 4.4: Principle component plot of newly sequenced ancient samples and reference modern individuals performed using the finestructure library. Green labels correspond to Migration Era samples, red labels correspond to Early Middle Age samples and white labels correspond to reference populations. 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.

To explicitly test whether Erg1 and DIN2 group together to the exclusion of other ancient samples and therefore, whether they likely originated from a similar source population, I performed f_4 tests in the form of $f_4(W = Erg1, X = DIN2; Y = test, Z = Mbuti)$, where *test* is 143 ancient populations used in the F-statistics analysis. This tests whether Erg1 and DIN2 form a clade to the exclusion of *test* or not. Of the 143 comparisons, only the population labelled as WHG had a $|Z| > 3$, ($Z = 3.057$), suggesting that Erg1 and DIN2 originate from the same local population. However, this result was surprising given we would not typically expect an individual from the LBK culture to form a clade with hunter-gatherer populations at the exclusion of a closely related farmer; this could be indicative of gene flow between a WHG-like source and Erg1. This result was robust to outgroup choice.

To determine whether Erg1 showed increased genetic similarity to local farming populations, I also performed combinations of f_3 in the form of $f_3(A = Erg1, B = test, C = Mbuti)$, where *test* iterates across 143 ancient populations. This tests the branch length, or the amount of genetic drift that has occurred on the branch between Erg1 and *test* since their divergence from an outgroup. The sample/population with the highest f_3 statistic was NE7, a sample from 4,360 – 4,490 BC and the Lengyel culture (a Neolithic culture centered on the Danube River, known to be an offshoot of the LBK culture Erg1 belonged to). On the other hand, DIN2 shows a clear affinity to samples from Neolithic France.

My dataset included data from several other LBK populations local to Erg1 and DIN2; samples from Schwetzingen, Stuttgart-Mullhausen and Halberstadt. These samples appear to form a distinct cluster on the plink PCA and are shifted away from the primary cluster of Neolithic individuals and towards samples from the Anatolian Bronze Age and Baden Culture (a central European Chalcolithic culture) (Fig. 4.3). I wanted to know which LBK population Erg1 and DIN2 were closest to. I found strong evidence ($|Z| = 7.97$) that Erg1 shared more alleles with LBK populations from Schwetzingen than with Stuttgart-Mühlhausen, suggesting the early LBK populations showed relatively fine-scale geographic structure. Given the lack of Hunter Gatherer ancestry in the Rivollat LBK samples, this structure seems unlikely to be driven by variable amounts of Hunter-Gatherer admixture (Fig. 4.7).

4.3.3 Variable amounts of local hunter-gather ancestry in Neolithic farmers indicates a structured population

Prior research has shown that admixture occurred between newly arrived farming immigrants from Anatolia and local hunter-gatherers [79, 123, 147–149]. The position of Erg1 on the PCA, shifted slightly north towards the majority of the Bronze Age samples, suggests that it may have a component of Hunter-Gatherer ancestry. I applied the SOURCEFIND algorithm

to the 'ancients painting' co-ancestry matrix to infer ancestry proportions for all newly sequenced individuals, fixing 3 surrogate populations at WHG, Yamnaya and Anatolian Neolithic (Fig. 4.8). I inferred 26% WHG ancestry in Erg1, suggesting it may have had a relatively recent ancestor who was a Hunter-Gatherer. I inferred a smaller proportion of WHG ancestry into DIN2 (8%), implying they were perhaps were part of a structured local population, where different elements received varying amounts of hunter-gatherer admixture. qpAdm modelling broadly agreed with these estimates and showed that Erg1 can be modelled as a mixture of Anatolia Neolithic (61%, se=0.095) and WHG (0.3855%, se=0.095). Erg2 showed no evidence of hunter-gatherer ancestry and could be modelled directly as Anatolian Neolithic farmer.

To localise the closest source of Hunter-Gatherer admixture into Erg1, I re-performed the 3-population SOURCEFIND analysis, but instead split the WHG surrogates into more fine-scale groups; Loschbour, LaBrana, Bichon and the 2 'Iron Gates' (present-day Romania and Serbia), leaving 6 surrogate populations in total. This analysis showed that the two 8800-year-old Iron Gates individuals from Serbia contributed towards 33% of the ancestry of Erg1, showing that it was likely to be closest population to the mixing source in our dataset. To confirm that this was not an artefact of there being 2 Iron Gates individuals (where all of the other WHG populations had a single sample), I removed the lowest coverage Iron Gates individual from the surrogate pool and repeated the analysis. The proportion of ancestry inferred from Iron Gates was similar (31%), suggesting the sampling did not affect the inferred proportion. The same result occurred across all 5 independent SOURCEFIND runs.

To determine the date of admixture between an Anatolian Farmer-like and WHG-like source into Erg1, I used MOSAIC [139], which infers admixture events using a similar technique to chromosome painting. MOSAIC is able to model the 'true' admixing sources and determine the genetic differentiation between those and the sampled sources, in addition to the date of admixture. When modelled as a 2-way admixture event, MOSAIC inferred similar WHG and Anatolia Neolithic mixing proportions to SOURCEFIND. It inferred the cluster of Italian hunter-gatherers to be the closest population to the true mixing source (Fig. 4.6). MOSAIC is also able to infer the f_{st} between the 'true' mixing groups and the sampled populations. I inferred very low f_{st} between the true and source populations, suggesting we had sampled a good proxy for the 'true' mixing sources. I inferred an admixture date of 5.3 generations before Erg1 was alive. I caution that the admixture date may be unreliable due to only targeting a single individual and given MOSAIC boostraps over individuals (rather than over Chromosomes as in GLOBETROTTER or LD blocks as in qpAdm), it was not possible to obtain confidence intervals around admixture date.

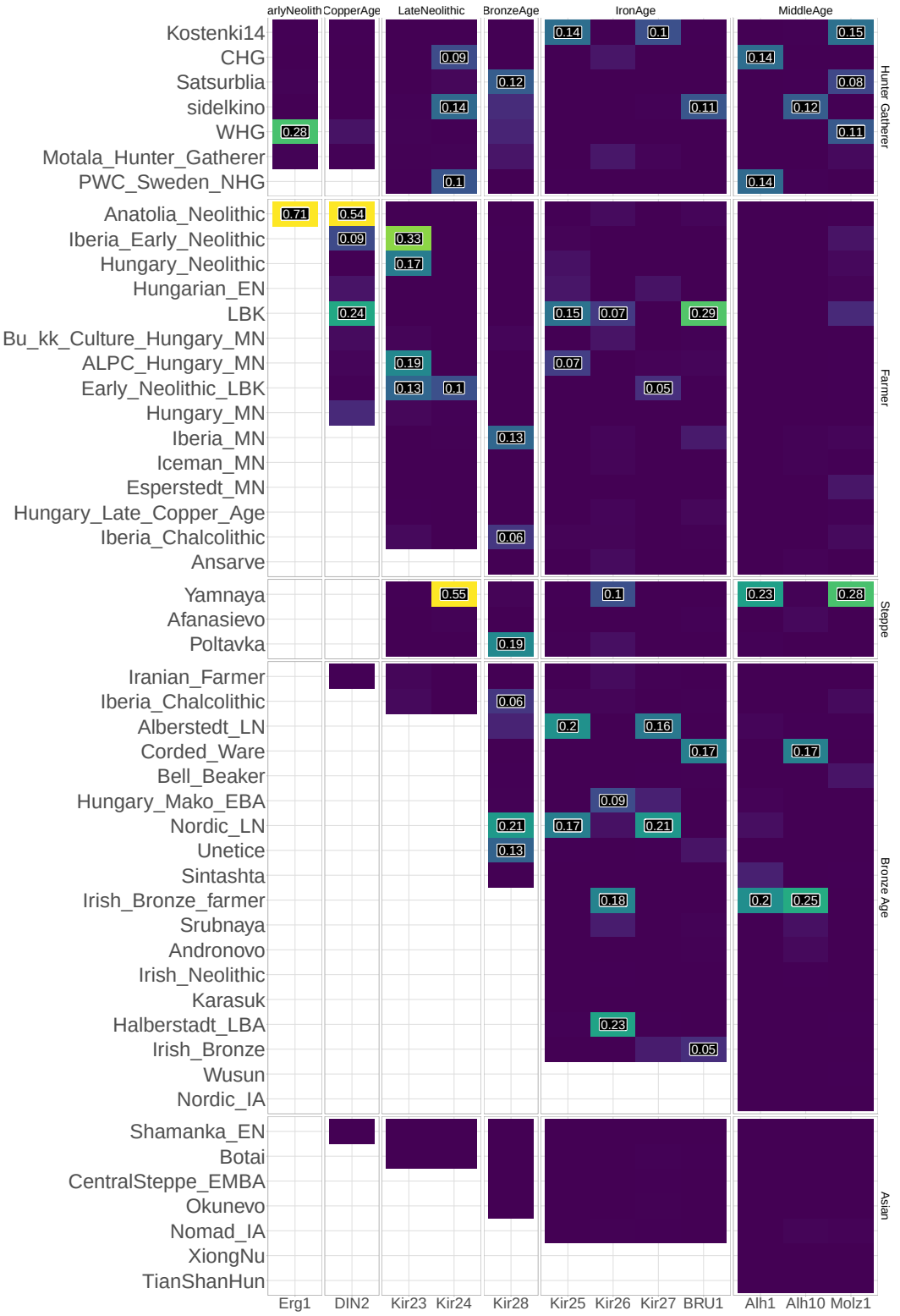

Figure 4.5: SOURCEFIND ancestry proportion estimates for all newly sequenced target samples (vertical columns). Target samples are grouped by archaeological age. Surrogate populations are represented as horizontal rows and also grouped into archaeological culture. Each target was modeled as a mixture of only populations which are dated to being older or contemporaneous as the the target. Numbers within each cell correspond to the ancestry proportion estimate.

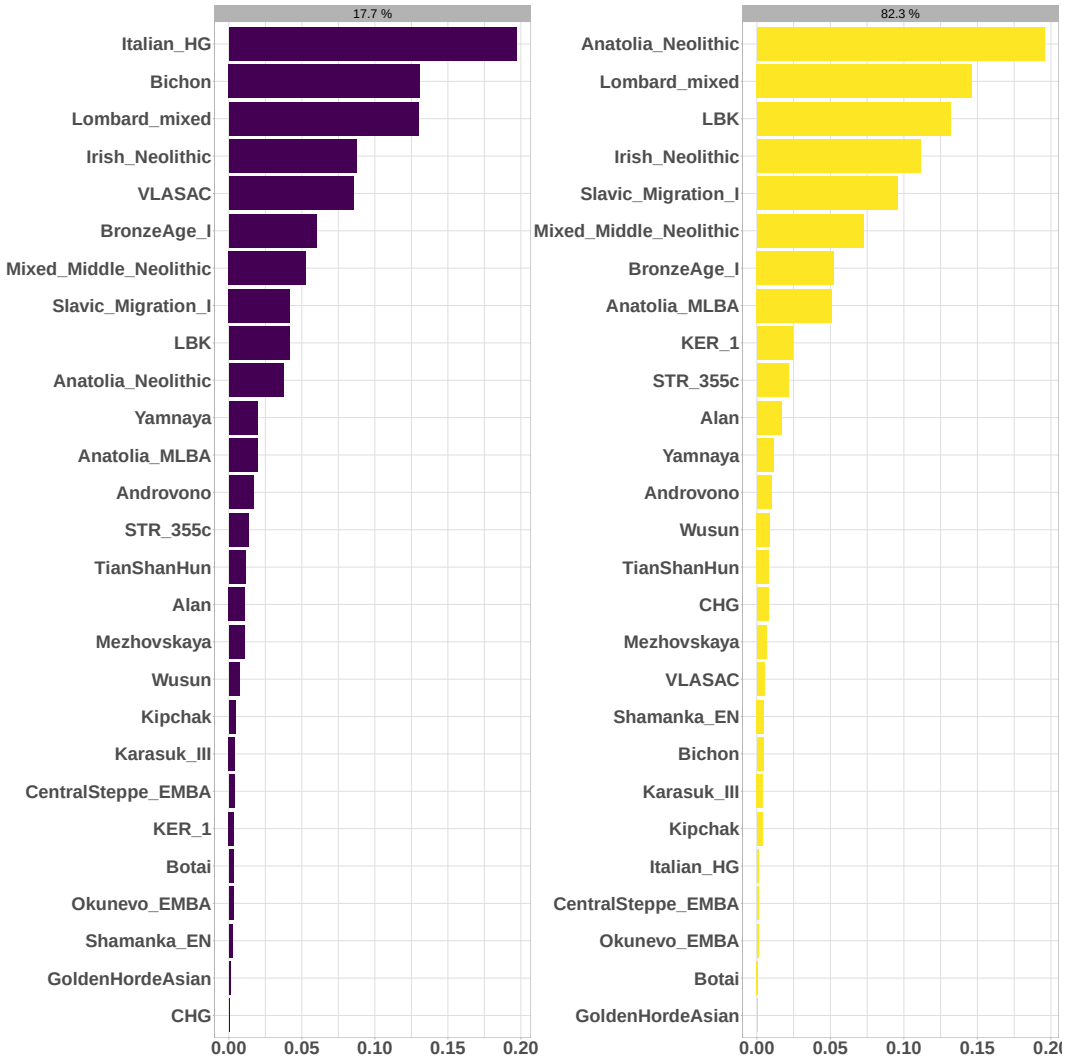

Figure 4.6: Copying matrix plot for sources in 2-way admixture event for Erg1. Each panel represents one of the 2 mixing sources. Labels above each panel gives the proportion that mixing source contributed to the Early Middle Age samples. Length of the bars within each panel represent the amount that mixing source copied from a particular population.

To confirm this admixture event, I performed an f_3 admixture test, which, when significantly negative, provides unambiguous evidence of an admixture event [29]. I performed the test $f_3(A = CastelnovianMesolithic, B = AnatoliaNeolithic, C = Erg1)$, selecting the A and B populations as those were inferred by MOSAIC to be closest to the admixture sources. This did not yield a significant result ($Z = 1.96$). However, exchanging Anatolia Neolithic for LBK, a source temporally and geographically more proximate to Erg1 yielded a significant result ($Z = 4.25$).

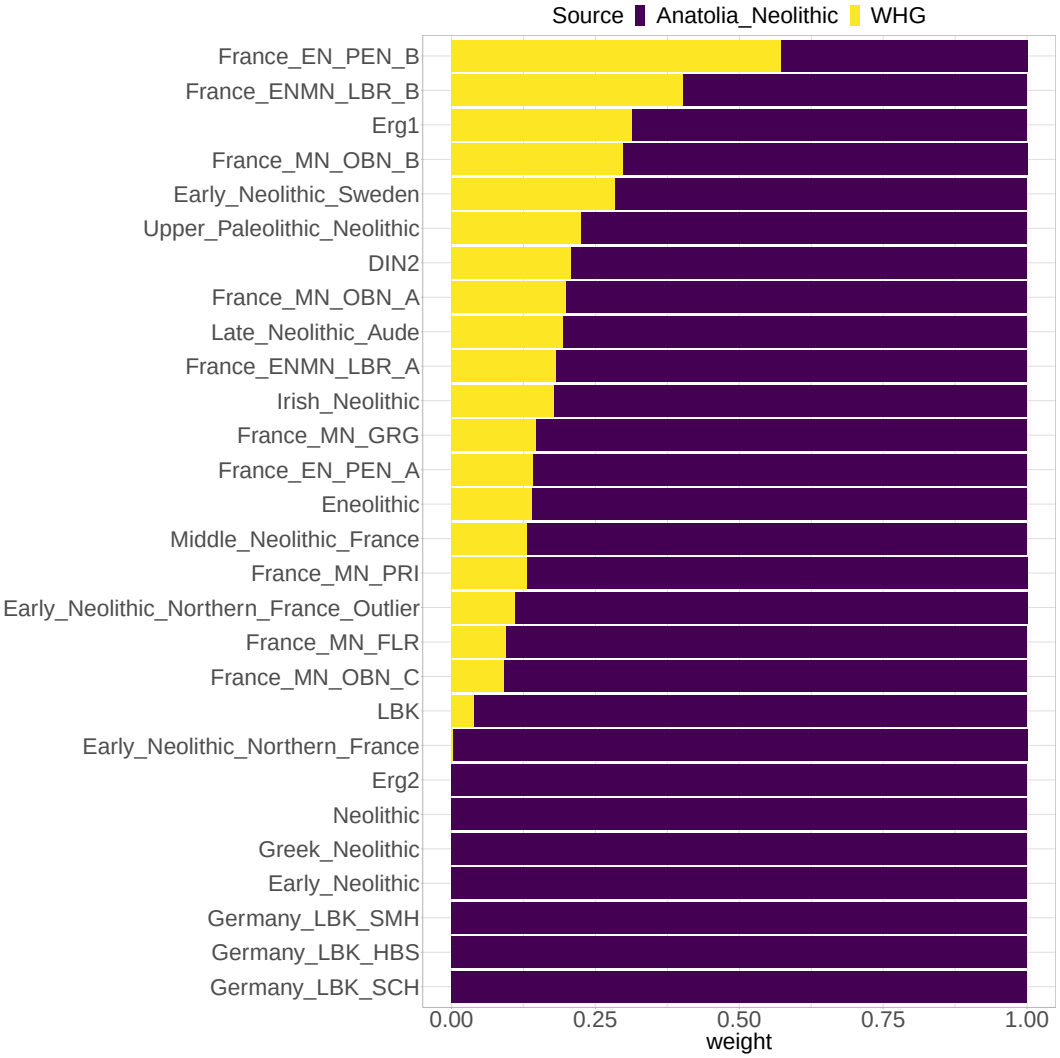

Figure 4.7: qpAdm ancestry proportion estimates for a selection of European Neolithic individuals. All individuals were modeled as a 2-way mixture between Anatolian Neolithic farmers and Western-Hunter Gatherers (WHG). Outgroups given in methods 4.2.9.

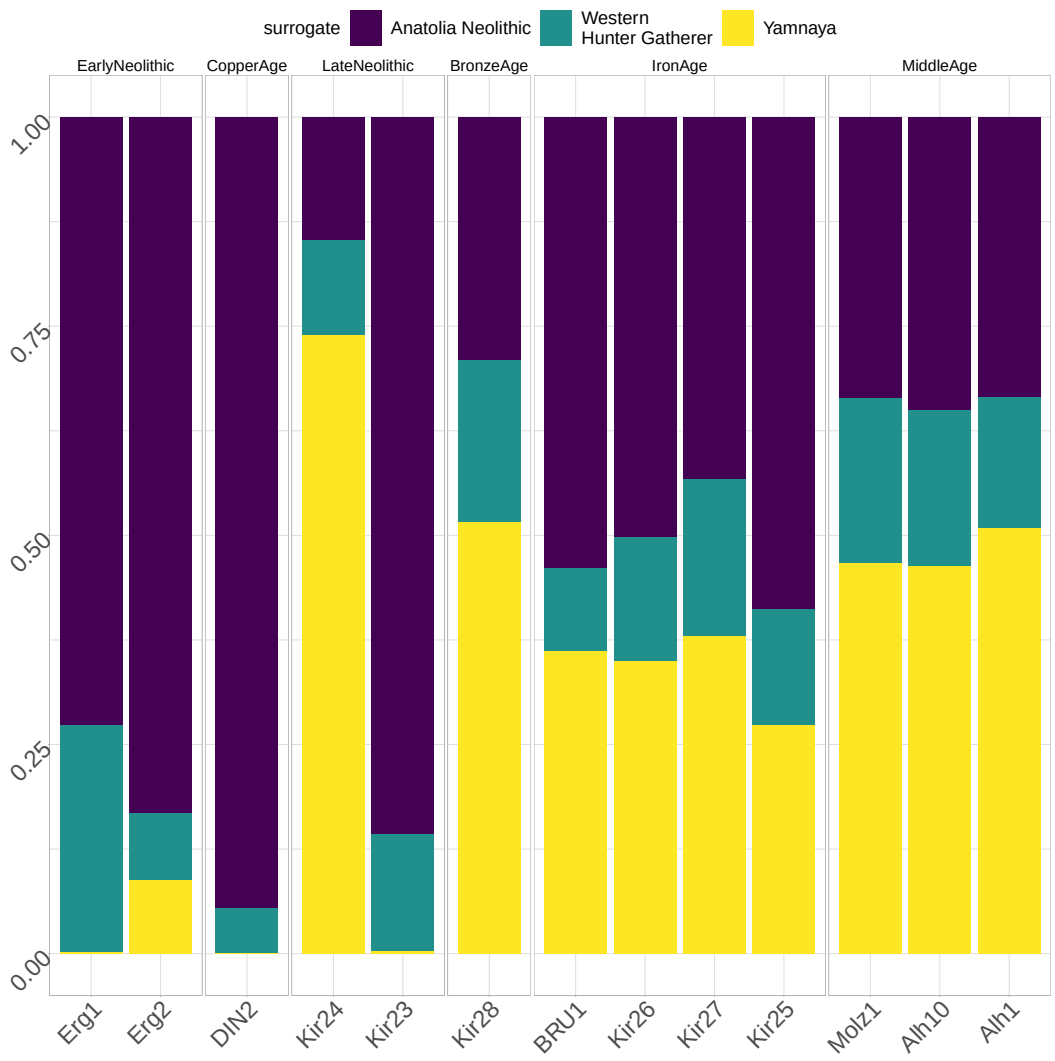

Figure 4.8: SOURCEFIND ancestry proportion estimates for all newly sequenced target samples (vertical columns). Target samples are grouped by archaeological age. Surrogate populations are represented as horizontal rows and also grouped into archaeological culture. Each target was modeled as a mixture of only populations which are dated to being older or contemporaneous as the the target. Numbers within each cell correspond to the ancestry proportion estimate.

4.3.4 Spatially and temporally close samples in Late Neolithic display highly distinct ancestries

This dataset included 2 individuals found in the same stratigraphical layer of Cherry-Tree cave; Kir23 and Kir24 were both dated to the Late Neolithic (approx 4700 BP). Despite their temporal and spatial closeness, they show highly different ancestry profiles (Fig. 4.8).

On both the plink PCA and fineSTRUCTURE clustering, Kir24 clusters with individuals from populations present around the Eurasian Steppe during the Bronze-Age, such as those from the Yamnaya and Afanasievo cultures. These are the populations known to be in part responsible for the spread of Indo-European languages across Europe [79]. These results support the findings of Allentoft (2015), who concluded that individuals from the Afanasievo Culture were 'genetically indistinguishable' from the Yamnaya Culture. That the Yamnaya and Afanasievo samples were sampled in Russia suggests that Kir24 may have been a recent migrant from the Eurasian Steppe. This is supported by IBD analysis; of all the ancient samples in the dataset Kir24 shares the most IBD (31.12cM) with the Yamnaya type-specimen and the lowest TVD with 2 other members of the Yamnaya population. This timing (Kir24 is dated to approximately 4700 BP) corresponds to some of the earliest appearance of Yamnaya-like ancestry in central Europe [150]. Using qpAdm, Kir24 could be modelled as a mixture of Yamnaya (93%, se=12%) and WHG (6%, se=8%) without any Neolithic ancestry.

Kir24 was assigned to mtDNA haplogroup T1a1, which has been found in Yamnaya samples from the Middle Volga region and Bulgaria [151]; the same study found the frequency of T1a1 to be higher in the Yamnaya peoples than in any other ancient or modern population.

On the other hand, Kir23 is found in a fineSTRUCTURE cluster with Ballynahatty, from Neolithic Ireland (3343-3020 BC), and is positioned on both plink and ChromoPainter PCAs with other late Neolithic samples. It is found in adjacent fineSTRUCTURE groups to samples from Neolithic Spain and Ireland. As is the case with other Neolithic samples of this era, Kir23 has a component of Hunter-Gatherer ancestry; it is known that Middle Neolithic individuals are characterised by admixture with the existing Hunter-Gatherer populations. qpAdm modelling showed that Kir23 could be formed from a mixture of Neolithic Anatolia (96%, se=14) and Hunter Gatherer (6.25, se=0.91) without the need for additional Steppe ancestry.

To test whether the source of Neolithic ancestry in Kir23 was most similar to local populations, I performed f_4 tests in the form $f_4(W = Kir23, X = mbutipygy; Y = test, Z = Erg2)$, which tests whether Kir23 forms a clade with Erg2, a local farmer individual, or *test*,

where *test* was one of several different farmer populations. Erg2 was chosen as the local group because it lacked any potentially confounding Hunter Gather ancestry. Kir23 always formed a clade with Erg2, suggesting that the source of ancestry into Kir23 was local and that there was a degree of continuity within the region.

4.3.5 'Southern' ancestry to Cherry-Tree Cave during the Iron Age is Italian in origin

The plink PCA shows that the four Iron Age samples are shifted towards the cluster of Neolithic individuals and away from the samples typical of the European Bronze Age. The same pattern is also seen in the present-day PCA, where the Iron Age samples are shifted substantially towards Spain / Northern Italy relative to the preceding Bronze Age sample which is situated among Northern / Western European populations (Germany, Wales) (Fig. 5.7). Previous studies into the Bronze-Iron Age transition in Western-Europe (France) have shown relative continuity [152]. Other studies in Eastern-Central Europe (Hungary) have shown the Bronze-Iron Age transition was accompanied by an increase in Eastern-European ancestry (albeit from a single sample) [147]. I was interested to see whether the transition in Bavaria had elements of either of these phenomena.

To identify the possible source of 'southern' ancestry in the Iron Age samples, I formed each of the Bronze Age, Iron Age and Middle Age Bavarian populations as a mixture of all other ancient populations using SOURCEFIND. I detected a component represented by 'Renaissance', a population from approximately 1500CE Italy, which contributed towards 26% of the ancestry to Iron Age individuals, but was found in neither the preceding Bronze Age nor following Middle Age. Thus, Renaissance samples appear to be the closest proxy for the 'southern' ancestry source. qpAdm modelling showed that the Iron Age samples can be well formed from a mixture of the preceding Bavarian Bronze age sample and those from either Renaissance Italy, Imperial Rome, Imperial Rome Late Antiquity or 'Roman Solider' from Veeramah et al (2018). All other possible sources included with Bronze Age resulted into poorly fitting models. This suggests a model of admixture from populations best represented by those from post Iron-Age Italy.

To determine whether this was an admixture event, I grouped the Iron Age samples together and performed MOSAIC admixture analysis. In the 2-way admixture model, the Iron Age samples could be formed of a mixture of a source closest to an Alamannic-Frankish sample (510 – 530 AD) 17.7% and a source closest to Anatolian Neolithic / LBK samples (82.3%). The estimated F_{st} between the 2 mixing sources was 0.016, approximately equivalent between present-day Germans and Palestinians [153]. Bootstrapped dates estimated the

date to between 7.86 and 11.31 (95% quantiles) generations ago. This signal is supported by the fineSTRUCTURE groupings; all 4 Iron Age individuals were grouped alongside several Lombard samples and a Roman solider from 300AD.

Based on SOURCEFIND modeling with the extended older surrogates set, unlike Gamba et al (2014) [147], I found no evidence of East-Asian or East-Asian-like admixture (Fig. 4.5).

4.3.6 Present-day genomes unpick genetic differences between early Germanic and Slavic populations

Lastly, my dataset included 3 samples (1 newly sequenced) from the Middle Age period. The two genomes from Altheim, Germany, date to around 500AD and were found in a Roman context. The single individual from Molzbichl, Austria, dates to around 300 years later, and has been assigned to a 'Slavic' cultural context.

The 3 Middle Age samples appear to share common ancestry based on the plink PCA and are located next to other spatially and tempirally close samples from the Middle Ages. Some structure is apparent from the ChromoPainter PCA (Fig. D.6), with the two Altheim samples clustering more closely together to the exclusion of the Slavic sample; however, this difference appears to be subtle. f_4 in the form $f_4(mbutipygymy, Bavaria\_Iron; Bavaria\_Slav, Bavaria\_Germanic)$ returned a non-significant result, showing that samples from the Iron Age in Bavaria were symmetrically related to the later Middle Age sample. These results suggest that the differentiation between 'Germanic' and 'Slavic' populations arose post Iron Age. However this non-significant result could be caused by low sample sizes in the Middle Age populations or a lack of power in allele-frequency based methods.

The two Germanic samples fall into a fineSTRUCTURE cluster with a set of contemporaneous samples from Northern Europe, including 10-11$^{\text{th}}$ century Vikings from Estonia, Sweden and Iceland. On the other hand, Molz1 clusters with other individuals known to be from Early Slavic populations. Interestingly, the Slavic cluster also containing a sample DA29, also know as 'GoldenHordeEuro'. This sample is from Karasuyr, Kazakhstan, and has was dated to 1200-1400 CE. The Golden Horde was a Mongol khanate established in the 13th Century CE. Given this sample shows clear evidence of European ancestry and clusters alongside individuals from Early Middle Age Europe, it has been proposed that this individual was captured in Europe during the Mongol raids of the 13th Century, when they assaulted the Kievan Rus' federation. That 'GoldenHordeEuro' clusters with Molz1 suggests the location of capture in Europe may have been from Austria where Molz1 was found.

It is currently unknown whether, in addition to cultural and linguistic differences, genetic differentiation exists between the 'Germanic' peoples represented by the two Altheim samples, and the 'Slavic' peoples represented by the Molzbichl sample. All 3 samples are positioned close on the ancients PCA, suggesting they lack differentiation in the context of ancient samples. However, their positions on the modern PCA reveals there was strong differentiation between early Slavic and Germanic peoples (Fig. 5.7). Molz1 clusters with present-day Slavic speaking populations such as Poland, Ukraine and Belarus. On the other hand, the two Germanic samples cluster with present-day individuals from Germanic-speaking countries in Western Europe, such as Scotland, Germany and Wales.

Plotting differential haplotype sharing between the Slavic and Germanic sample makes this pattern clear (Fig 4.9). There is a clear division down the centre of Europe, dividing it into East and West that shows the structure in present-day Europeans has existed since at least the Early Middle Ages.

These results were recapitulated using SOURCEFIND, where I modelled each individual as a mixture of different modern-day populations. The two samples from Altheim derived a large proportion of their ancestry to modern day Germans (81.8%, se=12.8), whereas the Molzbichl sample derived a large proportion of its ancestry from modern day Polish (77.85%, se=20.3) and Croatians (11.7%, se=9.1).

4.3.7 Conclusion

Whilst the two samples from the Early Neolithic showed signs of being from at least closely related source populations, most notably from TVD and f_4 statistics, they also displayed variation suggestive of different population histories. Consistent with the hypothesis that DIN2 may have migrated along the Danubian route, it shares the lowest TVD and is found in a fineSTRUCTURE cluster with other samples from the Hungarian Plane. The two samples also showed differences in the degree of Hunter-Gatherer ancestry; whilst DIN2 showed no evidence of admixture, Erg1 likely had a recent Hunter-Gathere ancestor.

This dataset revealed that temporally and spatially close samples may have very distinct genetic ancestry profiles, with Kir24 and Kir23 showing high levels of Steppe-related and Neolithic ancestry respectively. In particular, Kir24 seemed to be very recently related to the Yamnaya type-specimen sample, sharing 31cM of IBD with it. The arrival of Yamnaya-like ancestry from this early (2762BC) represents one of the earliest known appearances in the literature.

I identified a possible incoming signal of 'southern-like' ancestry during the Iron Age,

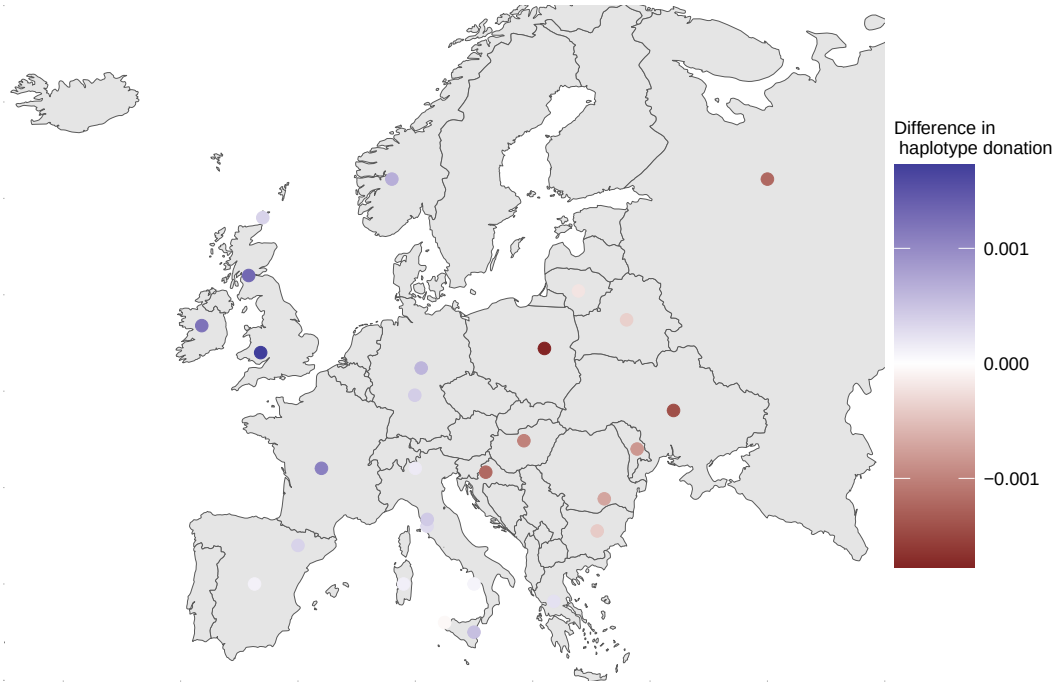

Figure 4.9: Differential haplotype-donation between Germanic and Slavic samples. Each coloured point is one present-day population. Points are coloured based on whether they donate relatively more to Germanic (blue) or Slavic (red) ancient samples.

which was not present in the single sample from the preceding Bronze Age. The most plausible source of this ancestry is from Italy, with the best source in the dataset being the cluster of Renaissance samples from Antonio et al (2019) [43].

Finally, I showed that including present-day reference samples is sometimes necessary to identify genetic differences between closely related individuals; when analysed in the context of other ancient samples, the structure between the early Germanic and Slavic samples was subtle. However, when painted using a set of present-day Europeans, extremely clear differences could be observed. These results showed that structure between East and West Europe dates to at least 1000 years ago.

Future studies in this region should focus on obtaining a higher density of samples, in

particular from the Bronze and Iron Ages; the low number of samples from these time periods mean any results should be interpreted with caution. More samples would show whether the introduction of 'southern' like ancestry in the Iron Age was a widespread phenomena, or restricted to a smaller geographic region in Southern Germany. Similarly, a wider sampling of Iron Age groups from Germany, Italy and Switzerland may allow for a more accurate identification of this source.

Whilst the utility of using present-day genomes was outlined through the comparison of the Slavic and Germanic samples, the analysis would have been significantly improved with higher resolution data from Germany. The data I have, described in Appendix section A.4, only had country-level details. Data which had labels from different sub-regions in Germany, similar to the POBI dataset, would have allowed for a finer-scale investigation into the current East-West genetic divide in present-day Germany.