Phylogeography of R-U106

Iain McDonald

March 6, 2025

This white paper draft describes methodologies and data relating to the phylogeographic spread of the Y-DNA haplogroup R-U106 and its sub-clades. It is incomplete and presented for early review by the community. This is not meant to be an authoritative and fully correct account of the growth and spread of R-U106, merely to be less wrong than existing alternatives.

Contents

1 Introduction

1.1 Phylogeography

"All models are wrong, some models are useful."

Most simply, phylogeography maps the phylogenetic tree (the "haplotree") onto a geographical map. Human Y-DNA phylogeography therefore looks to identify where Y-DNA haplogroups came from. In an ideal world, we would have genealogies for every one of our male-line ancestors, so phylogeography would be simply putting them onto a map and observing where our common ancestors lived.

The reality is that most of us can only trace our male-line ancestors back a few generations, with typical earliest known ancestors (EKAs) living only about 200 years or so ago. Beyond this, we have to link people together using genetic testing, and place the locations of their most-recent common ancestors (MRCAs) by extrapolating from what is known about their EKAs.

That extrapolation can provide substantial errors. Serious problems include the following, in approximate order of seriousness:

1. Biased testing among populations. The majority of people taking Y-DNA tests come from the USA, where genealogy is culturally seen as more important than many other countries, where recent wholesale migrations mean people often are ethnically mixed in unknown ways, and where there is the economic potential for people to afford tests. Contrast this with countries like the UK, where people know where they have lived for generations and thus don't care about genealogy, poorer countries where people can't afford to test, and countries where Y-DNA testing is restricted or banned (e.g., France). This means that the countries with the most testers are often not the countries in which a haplogroup is most common.

2. Asymmetric migration of testers. For example, the population of Spain and Portugal is approximately 59 million, but there are many times this population of Spanish and Portuguese diaspora living in Latin America. If we take a haplogroup that is common in Iberia (e.g., R-DF27), and simply took the median position of everyone's EKAs' geographical locations together, we might deduce R-DF27 formed in the Americas, rather than Europe. We know this is not true because history records the main migrations from Europe to the Americas. More generally, the region in which a haplogroup is most common is not necessarily the one in which it first arose.

3. Accuracy of genealogies. The majority of genealogies reported on Y-DNA testing platforms are accurate. However, they can be in error due to either poor genealogy or hidden genealogical problems.

4. Accuracy of geographical information. Many Y-DNA testers have not shared their geographical information publicly, and the format of EKA information at the largest Y-DNA testing company, Family Tree DNA, is not in a machine-readable format. Sometimes latitude and longitude information is available but, more frequently, analysis has to rely simply on a country of origin. This may or may not correspond to the person's known genealogy, and may simply indicate the country from which they believe their ancestors have come.

These problems have meant that phylogeography of recent Y-DNA haplogroups has been described by some as "crystal-ball gazing", as a pseudo-science that has little merit in being able to achieve its objectives. Historically, that may have been true. A decade ago, direct-to-consumer "next-generation" testing was very much in its infancy, a good haplotree did not exist, and there were only coarsely sampled Y-STR data from a comparatively small number of testers from which to extract origins.

Modern solutions to these problems come from two main sources:

- Asymmetric migration of testers can be better identified by pinning the location of haplogroups' origins using ancient DNA, which can identify the presence of a haplogroup in a historical culture. If the identified haplogroup's time-to-MRCA (TMRCA) is sufficiently close to the date of the ancient individual, then this can better locate the geographical origin of the haplogroup.

- The bias in testing populations, the accuracy of genealogies and the accuracy of genealogical information can be better understood by self-consistently examining a larger sample size of genetic testers. Better efforts are also now ongoing to make EKA information machine-readable.

These solutions have the potential to lift phylogeography of Y-DNA haplogroups from pseudo-scientific guesswork to an actual science. This work does not claim to do that, as it still contains a lot of guesswork, but it takes steps in that direction.

1.2 Historical and personal background to R-U106

The haplogroup R-U106 is defined by the Y-SNP U106 (an equivalent SNP, named FTT10, has subsequently been discovered). U106 was first discovered in 2007[1]. However, it was independently discovered at Scotland's DNA and given the alternate name S21, and by Peter Underhill (Stamford), who give it the designation M405. Hence, the names

R-S21 and R-M405 have also been used historically to reference this haplogroup. Early phylogeography [1] identified R-U106 as being most common in Germanic-language countries, thus was born the popular myth that R-U106 is associated with the Germanic people. In reality, this is an over-simplification.

A number of other names have been assigned to this haplogroup, based on the old naming structure. R-U106 sits within haplogroup R1b, with R1 named as the first branch of haplogroup R, and R1b as the second branch of haplogroup R1. Being the third branch of R1b found, it was originally given the designation R1b3. However, as new branches continued to be found upstream of R-U106 (above it in the phylogenetic tree), this numbering evolved. Even by the time it was entered into the 2006 version of the ISOGG Y-SNP tree[a], it had been designated R1b1c9. By 2014, when next-generation sequencing became commercially available, it had been designated R1b1a2a1a1. In the most recent (2019) version of the tree[b], it is designated R1b1a1b1a1a1. The complexities of keeping track of this evolving terminology, and the unfeasibly long haplogroup descriptions that result meant that reversion to designations like R-U106 (or occasionally R1b-U106) occurred, and now this is the standard way of referencing the haplogroup.

Y-STR tests represent the means by which most people start exploring Y-DNA testing. These lack the ability to estimate precise haplogroups, and Family Tree DNA declines to give them a more accurate haplogroup the R-M269 (a much older relative of R-U106). However, the 66th marker in the now-standard Family Tree DNA set, DYS492, has proven a good test of whether a person is U106+: about 97% of men with both a R-M269 predicted haplogroup at Family Tree DNA and a Y-STR result of DYS492 = 13 are part of R-U106.

The R-U106 project at Family Tree DNA formed in 2008 to form a community of people who were part of this new haplogroup. Similar projects were formed for its sub-clades R-L1 and R-U198 and, later, R-Z18. The R-U198 and R-Z18 projects exist to this day, though the R-L1 project was folded into the R-U106 project in 2019.

I took my first Y-DNA test in 2008 and joined the R-U106 project in 2009. I became more active in the group in 2011, when we were able to group testers into clusters on the basis of their Y-STR tests. I focussed on one cluster that we later named the "Kings' Cluster" after it was discovered that the House of Wettin formed part. Early phylogenetic inferences here were marred by the quality of the data and a poor understanding of its biases.

Several studies, including Family Tree DNA's "Walk the Y" project, began uncovering new SNPs in 2012, and several were identified as being within R-U106. Our "Kings' Cluster" became named R-DF98, and many other SNPs formed new haplogroups. While uptake of testing these SNPs was low, it allowed firmer placement of testers onto a phylogenetic tree.

The arrival of the Big Y test (and several competitors) in late 2013, and the subsequent release of Big Y-700 in late 2018, finally allowed the discovery of novel SNPs and the accurate construction of a phylogenetic tree. It also allowed the creation of accurate TMRCA estimates. Initially, all of these tasks were performed within haplogroup projects by volunteers. However, they have all since been formalised as part of the Family Tree DNA database, either through their online haplotree or their Discover platform[c]. Simultaneously, the number of tests has increased roughly tenfold.

1.3 Existing phylogeographic studies

While improvements to data availability, volume and size have created the potential for significant improvements in phylogeography, the underlying methodology has yet to catch up. Early community efforts in phylogeography have included Hunter Provyn's Mygrations[d] and Rob Spencer's SNP Tracker[e]. While these tools were cutting-edge at the time (and, in many respects, they still are), they inadequately address the four problems outlined above.

At their heart, these two programmes provide a simple function. If a haplogroup has no sub-clades, then its origin is defined as the centroid: the average geographical position of all testers within that haplogroup. If a haplogroup has sub-clades, its origin can then be computed as the average position of all its sub-clades and of the testers for whom this is their most-recent known haplogroup (or "terminal" haplogroup). Various additions are made to each one, such as ensuring the centroids avoid locations in the sea, or pinning haplogroups to their "known" origins, but this doesn't solve the unknown aspects of the above four problems.

Family Tree DNA have since released the Globetrekker tool on their own Discover platform. While this adds substantial complexity to the decision algorithms, based on ease of migration and other factors, it still inadequately deals with biases, and makes insufficient use of evidence from ancient DNA.

An example of this is the R-U106 haplogroup R-Z156. For a variety of reasons (Section 5.5), R-Z156 likely originates somewhere in the Únetiče culture near Prague. SNP tracker and Globetrekker place its origin in the south-east of England, and Mygrations in northern France (having already detoured by south-east England). This discrepancy is a direct result of the four problems noted above, especially the bias towards testers from (or assuming they are from) the British Isles. Similar issues exist for other, otherwise-unrelated R-U106 sub-clades, e.g., R-Z2 and R-U198.

[a] `https://isogg.org/tree/2006/ISOGG_HapgrpR06.html`
[b] `https://isogg.org/tree/`
[c] `https://discover.familytreedna.com/y-dna/`
[d] `https://phylogeographer.com/`
[e] `https://scaledinnovation.com/gg/snpTracker.html`

2 Addressing sampling bias in the data

The easiest way to examine sampling bias in the Family Tree DNA dataset is to examine the stated countries of origin (often displayed in flag form on their system). We can begin with the total number of testers on the haplotree (i.e. in haplogroup A-PR2921) with stated origins in each country. Dividing the number of testers by a reference population of men from that country, we can identify that country's testing frequency, therefore how much the dataset is biased towards that country.

2.1 Defining a historical reference epoch

The reference population is, nominally, the population of a country. However, we must account for the migration of individuals between their EKAs and their present positions. For example, a person may state "Ireland" as their country of origin, but their family may have been in the USA since 1845.

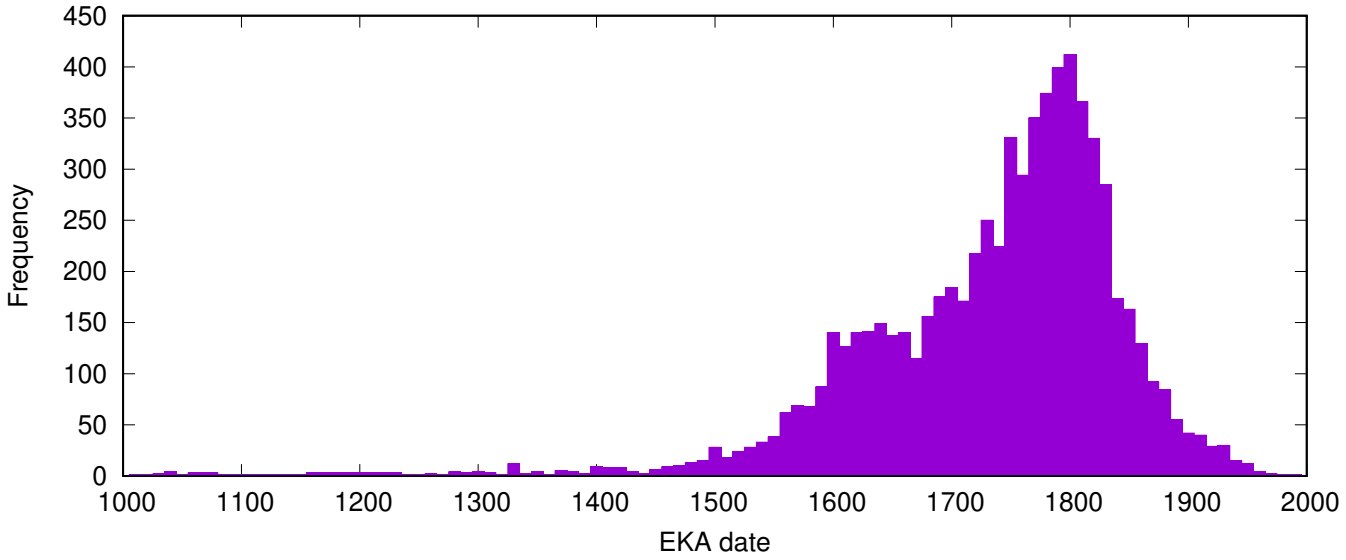

Figure 1: Histogram of the earliest-known ancestor dates for members of the R-U106 project. Data extracted from the Paternal Ancestry tables. Extraction assumes that any four-digit number in the person's paternal ancestry is a date, and that the first result in each person's line corresponds to the earliest record.

Figure 1 identifies that the typical tester within R-U106 has an MRCA that lived between about 1700 and 1850, so we can use this as an approximate range for a reference epoch. We can therefore judge our testing bias by setting the reference population to the population of each country during this period.

2.2 Constructing a historical reference population

Obtaining historical reference populations is not entirely easy for many countries, especially given the changes the political world map has seen since 1700. Appropriate historical census information and data from other records and estimates have been collected from the sources listed in Appendix A. Population data was then binned into ten-year periods for comparison.

2.3 Obtaining a testing bias

A sampling bias can then be determined by taking half of this population estimate (the male half) and dividing it by the number of testers from that country. Under the assumptions that 50% of the population is indeed male and that a negligibly small fraction of testers have TMRCAs with other testers more recently than ∼1700–1850, this gives the number of men living at that period per modern tester. Typically this number is a few hundred in 18th century Great Britain and Ireland, rising to the low thousands by 1850. In other European countries, it is typically thousands or tens of thousands.

We can establish a relative testing bias by normalising the number of men per test to one of the countries, or to the European sample as a whole. Since the UK and Ireland represent our largest testing population, it was chosen to normalise to the British Isles as a single unit. Averaging these bias factors over the period 1700–1850 approximates the relative testing bias for the Family Tree DNA testing database. Table 1 gives a list of these testing biases.

Several conclusions can be drawn from this data. First, the total number of R-U106 in Europe is likely to be ∼44 million men. This likely extrapolates to several hundred million worldwide. Currently, we appear to test a descendant for every 2369 men living in historical Europe.

Table 1: Bias factors for individual countries.

Country	Bias	Number of R-U106
England	1.547	5.55M
Scotland	0.623	323k
Wales	1.376	171k
N.I.	1.728	109k
Ireland	0.813	155k
Isle of Man	1.769	1k
Guernsey	16.792	7k
British Isles	**1.000**	**4.93M**[1]
France	23.457	3.13M
Germany	6.010	7.25M
Switzerland	3.816	0.45M
Belgium	20.482	1.09M
Netherlands	4.960	2.38M
Luxembourg	6.288	54k
North-West Europe	**10.148**	**14.9M**
Poland	5.340	0.91M
Czech	10.329	0.47M
Slovakia	9.546	0.13M
Austria	13.220	0.37M
Liechtenstein	2.641	2k
North-Central Europe	**8.488**	**1.86M**
Denmark	2.866	0.56M
Norway	1.002	0.32M
Sweden	0.884	0.55M
Finland	0.447	63k
Iceland	1.296	7k
Faroes	16.976	5k
Greenland	1.136	<1k
Scandinavia	**0.887**	**1.20M**
Russia	23.740	1.17M
Estonia	8.051	57k
Latvia	6.539	34k
Lithuania	1.985	32k
Belarus	5.978	72k
Ukraine	10.180	417k
Moldova	9.986	8k
Eastern Europe	**16.179**	**2.01M**
Albania	7.836	<1k
Bosnia	10.436	10k
Bulgaria	11.032	48k
Croatia	11.115	40k
Hungary	8.403	298k
Macedonia	10.130	19k
Montenegro	5.245	<1k
Romania	18.584	206k
Serbia	7.934	65k
Slovenia	3.148	63k
Turkey	20.230	108k
South-East Europe	13.341	**1.79M**
Cyprus	8.381	6k
Greece	7.754	53k
Italy	13.434	861k
Malta	4.307	11k
Portugal	7.551	207k
Spain	12.363	555k
Mediterranean	**12.620**	**1.75M**
Europe	**5.609**	**44M**
World	**6.467**	**414M**

[1]The British Isles population does not equal the sum of the component countries due to additional people specifying UK as their origin.

Most notably, however, are how significantly the bias factors vary from country to country. A generalised lack of testing in eastern and southern Europe exists, meaning that we completely miss many haplogroups that remain confined to these areas, and poorly sample many others. However, this lack of testing in the south and east is relatively uniform, so is theoretically much easier to account for when working with larger haplogroups where populations are reasonably sampled.

Much more difficult to deal with are cases like France, which is horrifically under-tested compared to its neighbours, due to the legal situation regarding paternity testing in the country. This lack of French testers creates a hole in the map, which can be very dangerous. If we see a haplogroup that is primarily Irish, British, Dutch or German, we may think that its origin may lie in that country, but it could be that the haplogroup is much more common in France, but severely undertested. For example, a haplogroup founded in Normandy might have migrated to England in the 1066 invasion. If half the haplogroup now lives in England and half lives in France, and there are 15 English testers, there is less than a 50% chance of having a single French tester. Without that French tester, we might completely ignore the possibility of a French origin, despite half the haplogroup living there. We would have to have more than about 80 English testers to be more than 95% confident that there was no majority French component of the haplogroup and (for the reasons explored in the next section) this still does not rule out a possible French origin.

This therefore sets the minimum size of a haplogroup where statistical mapping of country-level origins stands a reasonable chance of being able to elucidate an origin point. This minimum size is somewhere over ∼100 testers if the majority of testers are from the British Isles (or the Nordic countries), several tens of testers if the majority of testers are from western Europe, or a smaller number if a haplogroup exist only in eastern or southern Europe, or France.

3 Mathematical issues in phylogeography

3.1 Expansion and diffusion

An isolated population, spreading out unencumbered at a constant rate, will expand linearly, covering a distance d from its origin in time t. However, a population that intermingles and competes symmetrically with an existing population will behave broadly according to Fickian diffusion, spreading as $d \propto \sqrt{t}$. The latter, diffusive scenario is likely to work better on small, familial scales where individuals are important within a population; the former, faster scenario is likely to work better on larger migratory efforts, including invasions, where there is an asymmetric balance of power and proliferation. Given we consider country-level migrations in this text, the faster, linear timescale is probably more appropriate.

Most prehistoric and early historic individuals lived close to their parents. Prehistoric diffusion rates on timescales of millennia are typically on the order of ∼0.3–1 km/year[2,3]. We can therefore begin by naïvely estimating that a 5000-year-old haplogroup like R-U106 may have spread so that roughly half its population can be encompassed by a circle 1500–5000 km in radius. Most of this discussion on origins in this text focusses on a region of Europe with a radius of a little over 1000 km. Therefore, we rely on relatively subtle changes in distribution to track the motion of haplogroups older than ∼1000–3000 years. However, it is worth noting that, at least in historical times, the diffusion rates of the elite classes greatly exceeded that of the common man. Given the outsize role that elite classes play in the spread and success of haplogroups, this difference may be worth bearing in mind.

3.2 Founder effects

However, migration is not simply expansive or diffusive, as multiple factors mean a population will not expand symmetrically. Geographically, boundary conditions are imposed by the sea, across which a population cannot diffusively expand. Artificial borders can do the same under certain conditions, but do not appear to have had a significant impact on diffusive migration across history. Terrain conditions can also encourage or prohibit motion, with regions such as the Alps providing a barrier to migration across them, though rivers encourage migration along them. Crossing barriers such as seas, borders or terrain often requires the concerted action of one or more individuals to travel an uncommonly long distance. If a successful individual does this, then they may found a new haplogroup in their destination. The Discover algorithm attempts to take these boundary conditions into consideration.

This kind of "founder effect" has important implications for understanding origins. For example, many sub-clades of R-U106 were founded in among the Germanic peoples, where they had great success, and the locus of R-U106 in Europe is now among the Germanic-speaking countries. As mentioned in Section 1.2, this led many to assign R-U106 to be a Germanic haplogroup. Yet the haplogroup was not founded among the Germanic peoples: it pre-dates them by over 2000 years, comes from a completely different culture in a completely different part of Europe, and many haplogroups within R-U106 have never been associated with Germanic peoples.

This misunderstanding arose from a skew by which populations of R-U106 were more successful in western Europe than they were in eastern Europe, meaning the locus of R-U106 individuals moved westwards from its origin. Undoing these skews by identifying and accounting for these founder effects is an important step towards uncovering origins.

Table 2: Concordance of genealogical information by country

Country	Total	# with data	Rate of return	Raw counts (1)	Raw counts (2)	Raw counts (3)	Error fraction (1)	Error fraction (2)	Error fraction (3)	Max. error rate
Austria	27	23	85%	5 / 8	8 / 15	1 / 15	38%	35%	6%	47%
Belgium	58	45	78%	3 / 18	4 / 37	1 / 16	14%	10%	6%	34%
Czechia	55	47	85%	1 / 14	0 / 44	0 / 12	7%	0%	0%	17%
Denmark	83	70	84%	3 / 37	2 / 63	2 / 33	8%	3%	6%	22%
England	1636	1322	81%	237 / 564	245 / 865	34 / 555	30%	22%	6%	42%
England (Cornish)	16	16	100%	1 / 13	0 / 13	0 / 11	7%	0%	0%	*
Finland	98	86	88%	1 / 58	2 / 80	0 / 54	2%	2%	0%	16%
France	141	99	70%	4 / 29	7 / 88	0 / 28	12%	7%	0%	37%
Germany	796	610	77%	91 / 196	97 / 456	15 / 215	32%	18%	7%	41%
Hungary	41	37	90%	9 / 14	10 / 24	1 / 19	39%	29%	5%	40%
Ireland	425	313	74%	66 / 123	94 / 168	11 / 127	35%	36%	8%	54%
Italy	55	51	92%	0 / 22	1 / 49	0 / 31	0%	2%	0%	11%
Netherlands	192	168	88%	12 / 68	9 / 151	0 / 72	15%	6%	0%	21%
Northern Ireland	93	69	74%	6 / 29	7 / 47	2 / 18	17%	13%	10%	42%
Norway	187	177	95%	1 / 168	0 / 153	0 / 145	1%	0%	0%	10%
Poland	101	83	82%	2 / 41	4 / 75	1 / 38	5%	5%	3%	25%
Portugal	18	17	94%	2 / 7	0 / 16	2 / 6	22%	0%	25%	*
Russia	43	35	81%	2 / 19	2 / 30	0 / 18	10%	6%	0%	28%
Scotland	666	525	79%	103 / 203	108 / 332	19 / 202	34%	25%	9%	45%
Spain	42	31	74%	2 / 13	9 / 22	1 / 14	13%	29%	7%	50%
Sweden	333	303	91%	5 / 165	1 / 278	3 / 143	3%	0%	2%	12%
Switzerland	65	58	89%	4 / 22	10 / 45	1 / 22	15%	18%	4%	32%
Ukraine	31	27	87%	0 / 4	1 / 25	0 / 3	0%	4%	0%	19%
UK	1881	446	23%	245 / 849	254 / 1408	39 / 836	22%	15%	4%	67%
Wales	55	42	76%	12 / 10	10 / 26	2 / 14	55%	28%	13%	49%
Total	5805	4696	81%	679 / 1981	736 / 3387	106 / 1982	34%	22%	5%	41%

(1) Country and paternal ancestry information don't match / do match; (2) country and latitude/longitude don't match / do match; (3) paternal ancestry and latitude/longitude don't match / do match. Countries/regions with fewer than 16 respondents are not listed. *Result is dominated by small-number statistics, so not meaningful.

3.3 The fallacy of genetic diversity

One method to separate founder effects from origins is to look at genetic diversity. This diversity might be measured, for example, by the number of sub-clades present in an area, or by the variance of its Y-STR marker alleles. In any population, as mutations occur, this diversity should grow with time, even if the population waxes or wanes. However, diversity can be deceiving[4].

From a simple perspective, we can imagine that a man is part of a haplogroup, and that that haplogroup has a high diversity in his place of origin. If he then leaves that place and moves to a different place, he will form a new sub-clade of the original haplogroup in that location. Although the diversity of this new sub-clade will also grow with time, it will always remain lower than the diversity of the whole haplogroup, so the origin of the haplogroup will always remain identifiable as the location with the highest diversity.

However, most important migrations tend not to involve only one individual, but a wider culture exporting a fraction of their individuals wholesale to a new location (e.g., a fraction of the Normans moving to England following the 1066 invasion). Even if only a small fraction of the original population migrates, they will carry the majority of the haplogroup's diversity with them. This means that both the old and new locations (in the example, Normandy and England) will have the same genetic diversity, and the origin of the haplogroup will not be identifiable.

Diversity can therefore sometimes be used as an indication of origin, but the circumstances in which it works are limited.

4 Accuracy of genealogical information

4.1 Reporting accuracy

Public information on testers' origins is generally limited to country-level statistics, with the exception of some regional additions (e.g., in England, Cornwall). These regional additions are recent innovations at Family Tree DNA, so have not yet been properly utilised by the majority of testers in those regions.

These country-level statistics are provided by the testers themselves, nominally based on their known genealogies. However, in many cases, these country assignments are based not on pure genealogy, but on assumptions and incomplete

information.

To inspect these issues, Family Tree DNA's paternal ancestry table for the R-U106 project has been investigated. For those individuals reporting European countries of origin, the paternal ancestry, and the latitude/longitude pairs have been scanned. Concordance or discordance among these three records of origin has been assessed and summary statistics are presented in Table 2. We can draw several general statements from this data:

- There is a wide variation in the error fraction in various countries. This is generally determined by two main drivers:

    1. *Age and historical boundaries of countries:* relatively young countries (e.g., Belarus) generally have very low error rates, as individuals claiming ancestry from these countries can generally pinpoint an exact location of origin. Older countries (e.g., Austria) generally have higher error rates, as individuals claiming ancestry from these countries often put them as markers where the exact location is unknown (e.g., Austria for the Austro–Hungarian Empire; Germany for the Prussian Empire, etc.).
    2. *Emigration:* countries with lots of emigrants tend to have a greater fraction where the stated paternal ancestry and latitude/longitude does not match up to the country selected (e.g., claims of English ancestry when the stated paternal ancestry cannot be traced beyond the USA). In some cases, these claims have been proved incorrect or unlikely, though many will be true. Sometimes only a location of death is given: the location of birth may be known but unstated.

- Paternal ancestry often does not provide enough information to assign a point of origin. However, in a notable fraction of these null cases, the ancestry stretches to medieval times, thus could also be considered as being a correct match to the country flag.

- Generally speaking, testers with a specific latitude and longitude (within Europe) match on all three criteria.

- The rate of return (fraction of testers with either latitude/longitude or location in their earliest-known ancestor information) is normally much lower in the British Isles (23% for UK, 74–81% for constituent countries) than it is for the rest of Europe (70–95%). Nordic countries have generally more complete data (84–95%). Overall in the dataset of 5803 individuals, 46% had a location stated in their earliest-known ancestor information, 71% had a latitude and longitude, 36% had both. Within this 36%, 77% (28% of the total) had both pieces of information match the stated country of origin.

- Paternal ancestry and latitude/longitude generally provide better concordance with each other than with country-level data, and are often detailed enough to be genealogically believable.

We can also identify factors specific to individual countries:

- *Austria/Hungary:* A high error fraction arises from the historical borders of the Austro–Hungarian Empire. Unlike other countries, the presence of a precise latitude/longitude does not increase the fraction of tests within that country's borders.

- *England:* The large error fractions mainly derive from US-based individuals with English names. In many cases, the "English" nature of the family may be assumed from either the name or the history of the county into which they arrived. Many of these individuals are likely English, but have not provided information to suggest that they have proven that. Some places (e.g., Bristol) may be slightly over-represented, suggesting that these places are given as ports of embarkation rather than exact locations of origin. The effect of this is unquantified, but appears small.

- *Germany:* Highlighted errors in Germany often arise from confusion with Prussia. If latitude/longitude aren't given, "Germany" is accepted as a match but "Prussia" is treated as null: this assumes that the users have given the correct paternal ancestry themselves. The majority of erroneous locations should be sited in modern Poland.

- *Ireland:* Similar issues to England, with a large fraction of US testers. Locations within Northern Ireland were treated as errors, since Northern Ireland has always been a country of choice. If Northern Irish testers are allowed under the Ireland designator, a further $\sim$50 testers are recovered as matches (i.e., most of the errors).

- *Netherlands:* While the "van" and "van der" prefixes in the Netherlands can often pinpoint an origin, these have not been used. However, adding them would significantly increase the raw counts. Most errors in the Netherlands are due to testers in the USA.

- *Norway:* has very detailed origins for most of its countrymen, largely thanks to pro-active work by the Norway project.

- *Russia:* most errors here are from the Ukraine or Belarus.

- *Spain:* Spain's relatively high error rate arises from Spanish diaspora (mainly Latin America).

- *Switzerland:* Switzerland's error rate stems partly from American testers with possible Swiss ancestry, and southern German families with presumably inferred historical ties to Switzerland.

- *United Kingdom:* When the UK is generically selected, this brings in a lot of testers with American co-ordinates or paternal ancestries. The respose rate is generally much lower than the UK's constituent countries and, while there are a number of people without co-ordinates, few who do select co-ordinates use the centre of the UK without further qualification: a significant fraction can actually derive their ancestry as being from one of the constituent countries, and often specific towns or parishes.

- *Wales:* Wales appears particularly affected by the assumption that a common Welsh surname equals a Welsh ancestry. A significant number of Americans affect these statistics in particular.

From this, we can estimate an approximate error rate for each country and for the total. Error rates are considerably higher (235 out of 574 or 41%) among individuals who do not give supporting latitudes/longitudes, compared with those who do (736 out of 4123 or 22%). Assuming this trend is amplified in individuals who give *neither* latitude/longitude or supporting paternal ancestry information, this means that using error fractions among recorded data will underestimate the total error rate.

Based on these figures, we can estimate that the maximum likely error fraction among the countries in Table 2 (E') is:

$$E' \approx E_2 f_2 + E'_1(1 - f_2 - (1 - R)) + E_0(1 - R), \tag{1}$$

where E_2 is the error fraction #2 (for latitudes/longitudes) listed in Table 2, E'_1 is the error fraction among individuals who don't give supporting latitudes/longitudes (country-specific, but the above average of 41% suffices) f_2 is the fraction of individuals returning latitudes/longitudes (obtained from the raw counts divided by the total), R is the rate of return, and E_0 is the error fraction in people returning no data: this is unknown but must be lower than 100%. This calculation gives the final column in Table 2.

It should be stressed that this is a *maximum* likely error fraction. There are several reasons why the true error fraction is likely to be much lower, including:

- The error rate in null returns (E_0) will be less than 100%. Over all countries, null returns represent 23% of returns so, assuming E_0 is at least 41%, this could reduce the true error rate by up to 13%.

- Similarly, disproportionately many medieval genealogies (before the age of emigration) do not include a location. For noble families, this is often because locations are not needed: nobles are over-represented in the database, as they are both commonly targetted by genetic testing and were historically able to support bigger, healthier families. This affects a few percent of cases.

- Many claims of "errors" are based on the fact that a family has stated an ancestry in a different country (often the USA) but that they cannot trace beyond that country. In reality, this covers a full spectrum of reliability of information on the family's history. This may range from an assumed country of origin based on a surname or location, to documentary evidence that that person was an immigrant, to that they arrived on a particular ship from a particular country, to a document stating that they were born in a particular country, to a full genealogical history that they have simply chosen not to report. The rate of these errors is unclear, but is expected to account for a moderate fraction of the errors.

- Similarly, American R-L151 families have to come from somewhere in Europe within the last few hundred years (not least because none list "Native American" ancestry). We must assume that many of the American families claiming ancestry from a specific European country at least have the right country by accident if nothing else. Combined with the problematic use of surnames and/or residence in regions historically populated by (usually English or at least British) emigrants, this should result in many of the "errors" being accurate. This is expected to significantly reduce the true error fraction.

- Finally, many results that appear to be truly erroneous simply list the neighbouring country (e.g., Ireland [implicit, Republic of] instead of Northern Ireland, or England when they mean another part of the UK). Similar arguments can be made for historic Prussia or the Austro–Hungarian empires. These errors due cause problems in small-scale arguments, but do not significantly affect larger-scale migration questions, such as the location of groups in regional contexts (e.g., British, Scandinavian, central European, etc.).

As a consequence, we can set a maximum reasonable estimator of a 41% error rate in the provided country flags as a measure of genealogical accuracy. However, the true rate of error is likely to be much lower: "guestimated" figures could be only 15–30% for the UK; 10–20% for its constituent countries, Germany, Austria, Hungary and Spain; 5–15% for the majority of the rest of Europe; and even lower for specific sub-regions.

4.2 Genetic accuracy of genealogies

The above analysis does not take into account the accuracy of the genealogies themselves. Two factors can be considered separately: the rigour of the genealogical research, and the drift between genetic and social family trees.

4.2.1 Genealogical rigour

The majority of quoted genealogies have no basis except the tester's own research. Most are therefore brought together by people with no formal qualification in either genealogy specifically, nor history more generally. While many are indeed collated very rigorously, many others are based on scant and poor records, and may contain significant errors.

This fact is clearly evidence in claims of descent from various early medieval kings from whom no known male-line descendants exist, with some stretching back to pre-Roman times. While numerous viable claims to ancestry exist dating back to at least the 9th and 10th centuries AD, no European descents from antiquity can be reliably claimed, so these claims can immediately be dismissed as (at best) poor sanity-checking of secondary reports created by others. In some cases, the information clearly shows that the tester has missed the fact that the question specifically targets the *paternal* line.

However, these are only the obvious candidates. There will be many cases where people have tried to push their line that one generation further, but where confirmation bias or wishful thinking has stretched a possibility extracted from a scant record into a probability, which is then reported as a simple fact. The rates of genealogical errors introduced by such oversights are difficult to ascertain. Circumstantial evidence suggests that it is likely to affect at least a few percent of genealogies.

4.2.2 Genetic–social drift and NPEs

The decisions of when a genealogy can be considered "proved" (or at least probable, to whatever degree) blend into the issue of NPEs — canonically a non-paternity event but, in actuality, a term that covers a variety of issues. Here, we must consider three issues: any event that changes a surname (a "surname-discontinuity event" or SDE), a misattributed-parentage event (MPE), and a true non-paternity event (NPE). Different scenarios are explored on the ISOGG website[f], along with literature noting appropriate rates.

SDEs have been very common throughout history, and (in countries that had surnames during medieval times) were more common in the first few centuries of surname adoption, before they became standardised. A typical 1000-year-old medieval surname might have ∼50–80% surname replacement among its modern descendants, implying a typical rate of ∼2–5% per generation.

MPEs become genealogically more common as records become more scarce. A common instance is a child born before the supposed parents' marriage: often this is attributed to the later-married parents who raise the child socially, but in many cases one parent (usually the father) has inherited a child born out of wedlock. Illegitimacy and adoption rates vary considerably over time and space with social practices, and may range from ∼1–10% per generation (though appear usually towards the lower end). Most cases are genealogically identifiable, as birth records will often register a different (or no) father to the adoptive father, though census returns and other documents would not make the distinction and attribute the adoptive father as the genetic father. If only scant documents such as passenger lists or wills are available, it may be impossible to identify illegitimacies/adoptions, and MPEs become more frequent. Of course, MPEs can also occur at any point in a genealogy due to poor genealogical practice.

NPEs, in their strict sense, are rarer. These require cases where good records exist, but where the parent on the record is not the genetic parent. These rates are typically put at ∼1–2% per generation[5,6], and can mostly be attributed to infidelity (or, rarely, babies swapped at birth).

These three overlapping cases raise the rate of likely errors by maybe ∼5–10% in total for the average tester, although this rises to ∼25–45% by the 11th century unless triangulation of tests can be performed to confirm the genealogy.

Overall, we can therefore conclude that the country flags are a noisy but meaningful representation of the ancestry of testers, but that detailed analysis should make reference to additional information in the testers' EKA information.

5 Ancient DNA and origins

5.1 The rôle of ancient DNA

Modelling issues mean it is not possible to uniquely extract the origin of a haplogroup from modern testers. Fortunately, we can rely on ancient DNA to help us.

Ancient DNA helps primarily by reducing the time between the origin of the haplogroup. This reduces the amount by which a haplogroup has spread diffusively, narrowing the search radius down to a more meaningful range. It also allows us to undo the asymmetric migration that has happened between the ancient burial and the present day. If we return to the origin of R-DF27 discussed in Section 1.1, we know that R-DF27 arose in Europe and not in Latin America, not only because we have historical records of Spaniards and Portuguese migrating to Latin America, but because we have ancient R-DF27 burials in Europe but not in Latin America. Conversely, ancient DNA allows us to positively state that a haplogroup was present in an area at a given point in time.

Proving a haplogroup was *not* in an area at a given time is more difficult: in most cases, it is impossible to prove the complete absence of something. However, we can make probabilistic statements such as "out of X burials we see

[f] `https://isogg.org/wiki/Non-paternity_event`

none among our haplogroup of interest, therefore we can be $Y\%$ confident that the haplogroup made up less than $Z\%$ of the population". This methodology makes some assumptions, the most notable of which are the following.

1. We assume that all haplogroups in the population are equally mixed among social classes (since elite classes are more likely to have an identified burial throughout most of history). While ruling families demonstrate this is patently not the case in the short term, social migration, extra-marital issue and NPEs make this true in the longer term. The division between short- and long-term is very culture-specific, but is likely to be only centuries if we are to consider the whole gamut of surviving burials or a large and fast-growing haplogroup.

2. We assume that the sampled burials are representative of the overall population in question. For example, if we look at a particular combination of period and region (e.g., early medieval Ireland), then we may only be looking at burials from one or two sites. These may be associated with specific communities, that may not be representative of the overall distribution of haplogroups in that region at that time. Also, if two cultures co-exist at the same location, but one practices burial practices from which ancient DNA can be recovered (e.g., mummification) and one does not (e.g., cremation or sky burial), then the culture with the recoverable DNA will be massively over-represented. Given haplogroup diffusion within co-existing cultures, the effect of this is likely only significant for a very short period (a few generations to a few centuries at most). However, these differences can leave gaps on the scale of entire cultural packages (e.g., the Urnfield Culture).

3. We assume that all burials are sufficiently well sampled to indicate a positive or negative call within the haplogroup in question. In actuality, this is not true, with many R-U106 individuals only being called positive for haplogroups such as R-L151, R-M269 or R-M343 (R1b), but with no downstream haplogroups testable. In these cases, the stated percentage defines a *lower* limit to the potential percentage.

4. We assume each sample is independent of others, whereas some ancient DNA samples from the same location have been found to be close family.

Failure of these assumptions may lead to underestimation of the errors involved.

Not all ancient DNA burials are the equal, however. Their utility depends on how much closer to the haplogroups founder they take us, therefore how much of that diffusion and asymmetric migration is removed from consideration. A late medieval burial, for example, may tell us less than the paper-trail pedigree of a well-tested modern family. However, if we consider an ancient branch of a haplogroup like R-U106, having ancient DNA from an individual living 500 years after its foundation removes 90% of the migrations built up over the last 5000 years. This reduces the probable origin of the haplogroup from a circle 1500–5000 km around a modern tester (see Section 3.1) to one of only 150–500 km (i.e., a circle with an area 100 times smaller), and we can be ten times more certain that that circle has not been moved by asymmetric migration.

Ancient DNA therefore has great utility in informing us of the origins of a haplogroup. However, the number of haplogroups with ancient DNA close to their foundation is very small. These few burials allow us to pin a handful of haplogroups onto a map, into specific cultures at specific points. Onto these pins, we can hang our assessment of other haplogroups, on the understanding that (as we go further from these pins) we become more affected by the factors highlighted in Section 3.

5.2 Ancient DNA samples in R-U106

To establish R-U106 in the context of other haplogroups, an amalgam of 4517 ancient DNA samples was taken from the literature[7]. Removing female (sex = F or XX) samples left 2620 individuals, and removing samples with no Y-DNA haplotype left 2272 individuals. Filtering to European latitudes (>34.°8 N) and longitudes (32° W – 69° E) left 1989 samples. The sample I3035 was also removed, as the assigned haplogroup (R1b1a1b1a1a1c1a2b) is much more recent than the sample's age (5700 years)[g]. Note that this dataset does not include PNL001.

We can take this dataset in its entirety, or divide it into arbitrary shapes, including our consitituent regions and countries. By invoking a probability from binominal sampling, we can determine (under the above assumptions) a confidence interval using binomial probabilities, within which the true fraction of a haplogroup (e.g., R-U106) lies. Unless otherwise stated, we test probabilities of $p = 0..1$ using steps of 0.01, with a boxcar smoothing of 500 years, sampled in 50 year intervals between 5000 years ago and the present.

Within the sample of 1988 burials, 126 are typed to R-U106 sub-clades, 21 are only typed to R-U106 itself, 42 are typed to R-L151, 58 are only typed to haplogroups directly between R-L151 and R-M269, and 30 are only either R1b or R1b1a. These latter 130 (6.5%) stand some chance of being U106+, but the fraction is likely low (at most 30 per cent based on accurately typed results and modern percentages). This therefore imparts an additional $\lesssim$2% uncertainty to the fractions quoted above, which is small enough that we have not accounted for it.

Many articles on ancient DNA do not provide detailed haplotypes. The most common comparison is to the ISOGG 2019–2020 Y-DNA tree[h]. This lacks many of the detailed haplogroups of the Family Tree DNA haplotree[i]. A variety

[g] See `https://groups.io/g/R1b-U106/message/5875`

[h] `https://isogg.org/tree/`

[i] `https://www.familytreedna.com/public/y-dna-haplotree/R-U106/`

of individuals within the genetic genealogy community have searched the original data reads for haplogroups missed by published articles. A list of these additional reads for R-U106 ancient DNA is maintained by Raymond Wing[j].

5.3 PNL001 and the origin of R-U106

5.3.1 PNL001 and the Corded Ware Culture

PNL001[8] is a key individual in the origin story of R-U106. PNL001 was buried in eastern Bohemia some time between 2911 and 2875 BC when in his late 20s. While no isotope analysis has been performed to identify his place of birth, we can nevertheless date his birth to between 2941 and 2900 BC. He was buried in a simple grave with a bone awl, antler belt clasps and a cutting blade[9]. He had suffered two healed head traumas, plus a third that might have been the cause of his death[10]. He was one of three individuals from the 30th century BC Corded Ware Culture (CWC) in Bohemia (the others being VLI076, OBR003) who show strong autosomal similarities with each other at the ethnic level.

Cultural diversity in the early Bohemian CWC appears high, with some resembling both the earlier Globular Amphorae Culture (in which recovered males are mostly haplogroup I) and the Yamnaya peoples of the steppelands to the east. The three individuals above (VLI076, OBR003 and PNL001) have a very strong Yamnaya component.

5.3.2 The Yamnaya and ancestry before U106

The relationship between the Corded Ware Culture and the Yamanya is unclear. However, it is well determined that the Yamanya and their forebears are fundamental to the spread of the proto-Indo-European language[11] and linked to the spread of R-M269 in Europe[12] (specifically R-L151 and R-Z2103). R1a-M417 is also an important component of these migrations, appearing in burials from the middle of the third millennium BC in Denmark, Germany, the Czech Republic, Poland, Lithuania, and notably among the Fatyanovo—Balanovo culture in the forests of north-west Russia (the eastern equivalent to the CWC[13] where R-L151 is not seen).

The Yamnaya culture homeland in the Volga–Don steppelands is dominated by burials of haplogroup R-Z2103, which is removed from R-U106 at the level of R-L23, circa 5200–3600 BC. The lack of R-U106 (or even R-L151) forebears in eastern Europe and west Asia means we do not have a good idea of the origins of the R-U106/R-L151 ancestors before the CWC.

The Yamnaya-like properties of the Baden culture, the migration of R-Z2103 into the Hungarian plain by 2300 BC, and the presence of I18801 (R-L23, 2750 BC, Bulgaria) has been used to suggest that R-L151 arrived in central Europe via the Danube valley direct from an as-yet-unsampled part of the Yamnaya homeland. However, the presence of autosomal DNA components similar to Middle Neolithic Latvia has been used[8] to suggest an alternative origin in the forested north of the steppelands, with the CWC migration moving directly eastwards across Europe, north of the Carpathian mountains.

5.4 Refining the TMRCA of R-L151

5.4.1 Relative timings

Upstream "uncle" haplogroups of R-L151 (R-L51>PF7589, R-P310>FT186340, R-P310>FT123498) contain disproportionately large fractions of European testers for Asian haplogroups, but still retain some Asian (and notably Turkic and Arabic) testers, meaning these probably participated in the R-Z2103 "Kurgan" migrations into south-eastern Europe and Anatolia. This broader mix than the R-L151 hegemony in Europe suggests that R-L151 is the node on the tree at which the R-U106 line entered Europe. The rise of R-L151 therefore seems synonymous with the rise of the European CWC, and Family Tree DNA's projected TMRCA of R-L151 (3752–2408 BC, 95% c.i.) agrees with this interpretation.

If we accept the link between R-L151 and the rise of the CWC, then we can use archaeological constraints to narrow down the TMRCA of R-L151 within the millennium-wide range. The final result is very sensitive to the limits chosen. PNL001 places a hard constraint on the latest possible TMRCA of R-L151, but an earlier limit is more difficult to define. We can define two possible scenarios for the relative timing:

1. *Growth first*: R-L151 formed before the CWC migration. An existing R-L151 population migrated with the CWC. The lack of any uniquely eastern European / west Asian R-L151 groups means any remaining R-L151 in the steppelands died out.

2. *Migration first*: R-L151 formed during the early phases of the CWC migration once its westward travel had been initiated. Pre-L151 branches in the steppelands died/daughtered out.

Arguments against the growth-first hypothesis are the requirement that any remaining branches of R-L151 in the steppelands of easternmost Europe / west Asia have died out (the Danubian hypothesis avoids this problem only temporarily by taking a more circuitous pathway). The rapid branching of the R-L151 tree indicates that it grew very quickly in its initial stages. At some point, a haplogroup becomes "too big to fail": the haplogroup eventually

[j] `https://docs.google.com/spreadsheets/d/1rpJPOBt4qUQb9wWBFA7i1tLPV75ie_qS0iplwvvlVmQ/edit?usp=sharing`

contains too many individuals and the statistical likelihood that all of their lines will die out in a given time becomes negligibly small. The migration-first scenario does not suffer from this problem, as the lack of branching in the tree indicates a lack of population growth in the generations before the R-L151 founder.

Arguments against the migration-first hypothesis include the presence of PNL001 right at the beginning of the CWC's attested presence in central Europe. The CWC spread across Europe was rapid, moving from the Baltic States to the North Sea and the Alps in a matter of at most a century or two[14]. The formation of R-L151 (specifically the birth of the second son of the R-L151 founder) must have been during the very first phases of the CWC migration in order for the U106 SNP to have formed by 2900 BC.

5.4.2 General considerations

A TMRCA calculation for R-L151 is therefore limited at the young end by the requirement that the R-L151 founder must be at least the age of PNL001's father. We can take the carbon-14 date of PNL001 (2896 BC ± 17 years) and add the age of PNL001 at death (27.5 ± 2.5 years), giving the birth of PNL001 as 2959 BC ± 18 years (if adding errors in quadrature). The R-L151 founder must be at least one generation (33 ± 10 years) older than this, setting a reasonable expectation that the R-L151 founder was born before 2992 BC ± 20 years.

We are fortuitous that PNL001 represents some of the earliest evidence of the CWC, whose first arrival into Europe is estimated to have been in the few decades before 2900 BC[15]. This narrow interval also implies that PNL001 could have died on the frontier of the CWC as it expanded. The younger limit of 2992 BC ± 20 years for the age of R-L151 effectively precludes the migration-first scenario. This would require the R-L151 founder himself was among the first migrants, while simultaneously being a young father to the first U106+ man, who happens to be PNL001, and then only if PNL001 was able to father children before his early death.

When constraining the older end of range of possible R-L151 TMRCAs, it becomes a question of defining how big "too big to fail" actually is. Mathematically, the Galton–Watson process defines the probability ($P(x_n)$) that a male line will die out in any given generation (n), based on a natural population growth rate (λ), as $P(x_n) = \exp(\lambda(x_{n-1} - 1))$, with $x_0 = 0$.

Successful migrations come when the emigrant country enjoys a position of relative strength, and the immigrant country suffers a comparable weakness. The most successful migrations require both. We can therefore deem unlikely a mass-exodus scenario whereby an entire haplogroup is forced from their lands and then successfully takes over the entirety of northern Europe in the space of a few generations. The CWC migration is likely to be one of a strong group seeking to attain greater strength.

In such scenarios, the number of warring men is typically only a few per cent of the population, since men were needed at home for both defence and work, while many were too young, old or infirm to fight. For example, it is estimated that ∼5% of the Danish population migrated to England during the Danelaw (over several centuries)[k]. If we take 1–10% of a population as the size of a typical army, then allow for those too old to contribute to the gene pool and the female half of the population, we can then set a reasonable maximum that ∼25% of the viable R-L151 men migrated into the CWC while ≳75% remained at home.

5.4.3 Population growth rates (λ)

Considering the population growth rates, we have a post-migration λ_{CWC} in the CWC, which is derivable from the extant R-L151 lines in Europe; we have a pre-migration λ_{pre} in the R-L151 homeland, which must be $\gtrsim 4 \times \lambda_{\mathrm{CWC}}$ to account for ≲25% of men migrating; and we have a post-migration λ_{post} in the R-L151 homeland, which is unknown.

A reasonable expectation would be the average rate over the period 10 000 BC to 1700 AD, which is $\lambda = 1.013^l$, for which the Galton–Watson formula predicts a probability of survival of one in 34 after 150 generations. Therefore, if we estimate that at least 34 R-L151 men remain in the R-L151 homeland, then there is at least a 50:50 chance that there should be R-L151 men descended from them there today.

Of course, having men living there is different from having a testable lineage. We have not found all the R-L151 basal clades in Europe, and our coverage varies considerably. On the assumption that the R-L151 homeland is in easternmost Europe or west Asia, we can use the sampling bias for Russia as an estimate, which is about four times worse than the European average (Table 1). Therefore a European basal clade of R-L151 with only one tester has only a 25% chance of being detected in Russia, a European clade with only four testers has only a 50% chance of being detected in Russia, etc. This probably reduces the number of testable basal clades by about a third: i.e., if the R-L151 population today was equally split between two equal-sized populations in Russia and Europe, we would see three European haplogroups for every two Russian haplogroups.

Table 3: Number of clades below R-L151 by depth and sub-clade

SNP depth	R-L151	R-P312	R-U106	R-S1194	R-A8053	R-FTA1
1	5	1	1	1	1	1
2	6	1	1	2	1	1
3	29	19	2	4	3	1
4	46	24	5	13	3	1
5	63	28	12	19	3	1
6	98	54	18	22	3	1
7	189	139	22	24	3	1
8	282	225	27	26	3	1
9	406	329	43	30	3	1
10	495	407	49	35	3	1

5.4.4 Initial growth of R-L151: too big to fail?

A minimum population growth rate in the CWC comes from the number of haplogroups that exist at a given number of SNPs below R-L151. There are five known child clades of R-L151: R-P312, R-U106, R-S1194, R-A8053 and R-FTA1, so the number of clades at one SNP after foundation is five. These are all European haplogroups[m]. R-S1194 is the only haplogroup of the five to be represented by only one SNP, whereupon it splits into two, so the total number of clades at two SNPs after foundation is six. A full numbering down to ten SNPs after the R-L151 founder, current to the haplotree of December 2024, is given in Table 3.

We can see from Table 3 that one man gave rise to over 495 descendants in the time taken for ten SNPs to be generated. This is a minimum number, since many lines will have died out or remain untested. There is also not a unique mapping of number of SNPs to time, but a reasonable estimate would be 62.5 years per SNP[n], or about two generations. However, it gives us a good comparison for our remnant population, in which we would expect $3 \times 495 \times \frac{2}{3} \approx 990$ testable lines. Even if only one in 77 lines survives, the probability that at least one of the 990 would be tested today is virtually certain. Therefore, the lack of basal R-L151 lines with clear origins outside northern Europe means there is a probability of effectively zero that the R-L151 foundation was 20 or more generations before the CWC migration.

Applying the same logic, we can estimate that the R-L151 founder's birth pre-dates the CWC migration by no more than seven SNPs at 99.3% probability, no more than six SNPs at 92.3% probability, no more than five SNPs at 81% probability, no more than four SNPs at 70% probability, and no more than three SNPs at 53% probability.

This makes logical sense if we understand that R-L21, R-U152 and R-DF27 occupy very distinct parts of Europe, so cannot have existed for long enough that they mixed well into the invading CWC force. R-L21 splits from R-U152 and R-DF27 at the R-P312 level (two SNPs depth), while R-U152 and R-DF27 split at R-ZZ11 (four SNPs depth). We can therefore expect that R-P312 and especially R-ZZ11 cannot have significantly predated the CWC migration, and apply an earliest reasonable date to the R-ZZ11 split of 2950 BC. Similarly, we know that the ancient DNA individual RISE563 (2573–2310 BC) is R-ZZ11 and U152+, so must post-date the R-ZZ11 split by at least one generation. Therefore, we have four constraints:

- The R-L151 MRCA must be older than PNL001, as it is R-L151 and U106+.
- The R-ZZ11 MRCA must be older than RISE563, as it is R-ZZ11 and U152+.
- The R-ZZ11 MRCA should be younger than the Corded Ware Culture migration.
- The R-L151 tree cannot have grown much before the Corded Ware Culture migration, because we see no eastern-dominated clades and its many basal haplogroups. (This assumes the R-L151 ancestors came from considerably further east than Bohemia.)

[k] Population of Denmark in 800 AD: ∼500 000 (`https://natmus.dk/historisk-viden/danmark/oldtid-indtil-aar-1050/vikingetiden-800-1050/magt-og-aristokrati/hvor-stort-var-danmark-i-vikingetiden/`). Population of the Danelaw: 20 000 – 35 000[16].

[l] Based on a generation length of 33 years and a global population growth rate of 0.04% per annum (`https://ourworldindata.org/population-growth-over-time`).

[m] Note, however, R-FTA1 contains a single historical family of two individuals of American origin, one of whom has publicly declared the surname Rose; `https://www.familytreedna.com/public/R1bBasalSubclades`

[n] The Y-DNA point mutation rate in humans is $\sim 8 \times 10^{-10}$ SNPs per base pair per year[17]. The comparatively large number of tests in the four major haplogroups, and the discovery of new FTT series of SNPs in the T2T-realigned tests within R-U106 and R-P312 mean that the effective discovery space for SNPs for very well-tested branches is close to the full 23 Mbp of the "readable" Y chromosome, giving an effective rate of ∼54 years per SNP. For the smallest haplogroups, a rate of ∼81 years per SNP based on a 14–15 Mbp test would be more appropriate. Based on the relative size of haplogroups in this tree, an effective coverage of 20 Mbp is used as an average, equating to ∼62.5 years per SNP is obtained for the average rate.

5.4.5 The TMRCA of R-L151

Performing a TMRCA calculation[18] on these constraints, we arrive at a most-likely date for the TMRCA of R-L151 of 3115 BC, with confidence intervals at 3222–3029 BC (68% c.i.), 3366–2972 BC (95% c.i.) and 3507–2937 BC (99.5% c.i.). We use this likelihood function as our primary source for R-L151.

In reality, the true TMRCA for R-L151 is likely to be much closer to the younger end of this range. This is because:

- The growth rate among the remaining population is unlikely to have remained so close to replacement level ($\lambda = 1$) as soon as the CWC migration occurred.
- The CWC migration is unlikely to have been instantaneous, but progressed over several generations.
- The R-ZZ11 split is likely to have been at least a little after 2950 BC.

5.4.6 Population growth rate of R-L151 in the Corded Ware Culture (λ_{CWC})

From this data, we can also estimate a minimum likely growth rate in the R-L151 population, based on the number of haplogroups formed for a given λ, and the assumption that one SNP $\approx$ two generations $\approx$ 62.5 years. On this basis, we obtain a minimum $\lambda_{\mathrm{CWC}} > 1.344$, which approximates to 0.95% growth per annum over the first ∼625 years of R-L151's existence. This contrasts with the aforementioned ∼0.04% growth rate that typifies prehistoric populations, and is more similar to the world's population growth rate today. The true growth rate was likely much higher, as this estimate is based solely on the haplogroups that have survived to the present day. Rates in individual family branches must also have been much higher to generate the ∼20 child clades that we see in some haplogroups.

This growth was not equal among its constituent sub-clades, indicating very unequal reproductive success. R-FTA1, for example, must have experienced very little growth to be barely detectable today. The other four sub-clades, however, show an abrupt increase in sub-clade count roughly two SNPs (125 years) after the R-L151 foundation, equating to a brief period where $\lambda \gtrsim 2$, before settling into a slightly more sedate $\lambda \gtrsim 1.17 - 1.39$ for the following centuries. It is possible that this surge, corresponding to the foundation of R-P312, R-U106, R-A8053 and splits within R-S1194, is co-incident with the success of the CWC migration around 2900 BC. If true, this would place the foundation of R-L151 in the approximate period 3050–2980 BC.

The growth rate of R-U106 and R-S1194 remained approximately equal for the first few centuries of their history, with R-U106 being only ∼40% larger (49 versus 35 sub-clades) after ten SNPs (∼625 years). This suggests that the relative success of R-U106 today is down to its later expansion relative to R-S1194.

The growth rate of R-P312, however, is much higher, attaining a secondary growth spurt about six SNPs (∼375 years) after the R-L151 foundation. This corresponds roughly the splits of R-DF27 and R-U152, and the start of their presence in the archaeological record (from RISE563, ∼2542 BC). It is a combination of both the initial few generations of R-P312 and this later period that gives R-P312 its relative success over R-U106 and R-S1194. The size of R-P312, which represents 80% of the R-L151 sub-clades by eight SNPs (∼500 years) after the R-L151 foundation, reflects the relative ease with which P312+ ancient DNA is found compared to ancient DNA from other R-L151 sub-clades. This second rise could represent the successful integration and takeover that R-P312 made into the Bell Beaker culture[19].

The early period of R-U106 also sets its broad structure today. By eight SNPs (∼500 years) after its foundation, R-Z381 comprises 61% of the R-U106 sub-clades, R-Z156 comprises 16%, with the minor clades making up the remaining 23%. Notably, R-Z18 is yet to become a significant component.

5.5 The spread of R-U106 from ancient DNA

Figure 2 shows the fraction of R-U106 in ancient DNA samples in European (and west Asian) DNA overall (top-left panel), our main geographical regions (other panels) and England and Denmark specifically (bottom panels: these are the only two countries with sufficient U106+ samples to investigate). These plots are coloured by probability, with regions in red most likely and regions in yellow probable. Figure 3 gives the same graphs for the R-L151 fraction.

Care should be taken to differentiate rises in the red regions in Figure 2 (and green regions in Figure 3): those where the yellow (cyan for Figure 3) region fills the plot correspond to a lack of data; only those where the yellow (cyan) region is confined to a small range correspond to real rises in the fraction of R-U106 burials. Significant gaps in coverage for both occur around 1500–500 BC in different parts of Europe, due to the high prevalence of cremation burials during this period and subsequent lack of recoverable DNA. Care should also be taken regarding edge effects near modern times, due to the small number of sampled burials less than 900 years old. Note that the Galton–Watson process will mean that small sub-clades will be (often usefully) over-represented in ancient DNA.

5.5.1 Overall distribution

The overall distribution of R-U106 in ancient DNA is hard to decipher, since the entirety of Europe and west Asia is not well-mixed enough to treat as a single unit, and as the sampled burials are not evenly distributed over the region. Based on the fractions of R-U106 samples in each country at Family Tree DNA and the bias factors in Table 1, the modern-day R-U106 fraction in Europe and west Asia is ∼5%, though with strong regional variation.

The overall distribution of R-U106 in Europe and west Asia remains at or below this ∼5% modern value for most of its history, but exceeds it during the period 1000–2000 years ago. This corresponds to a rise in R-U106 fraction in

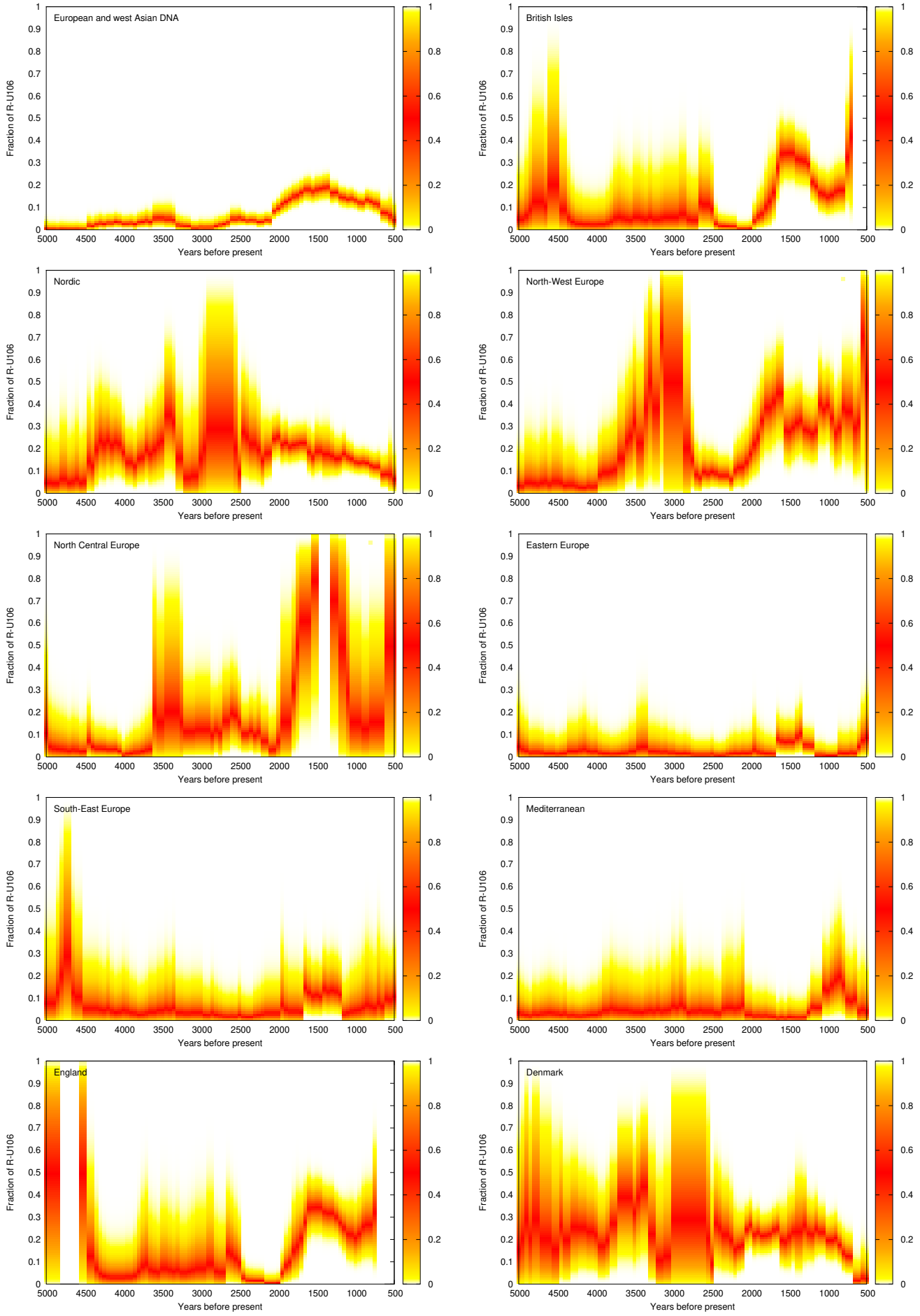

Figure 2: Fractions of R-U106 in ancient DNA samples by region.

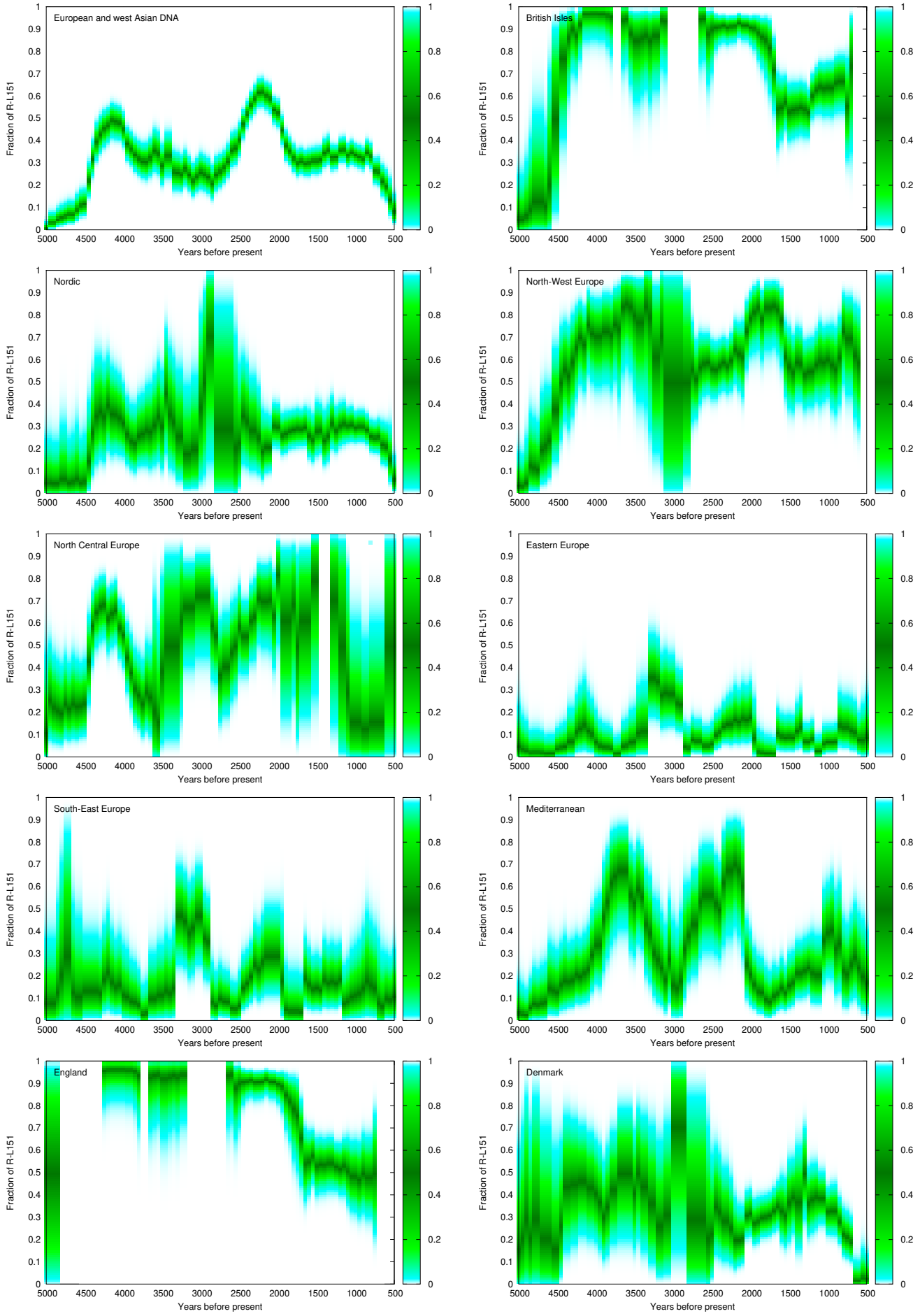

Figure 3: Fractions of R-L151 in ancient DNA samples by region.

several regions, including the British Isles, north-western Europe and possibly north-central Europe. This corresponds to the *Völkerwanderung*: the great migration period of the Germanic peoples after the fall of Rome. R-U106 is rightly credited with being a major component of Germanic groups, though the age of the haplogroup and its diversity mean that we cannot ascribe it the name of a "Germanic haplogroup" as some authors have. We may nevertheless be able to credit these Germanic migrations with a significant part of the prevalence of R-U106 in European cultures and their diaspora today.

5.5.2 British Isles

Ancient DNA in the British Isles is dominated by results in England. Scotland, Wales and the island of Ireland have too few burials among this dataset to explore well in isolation.

The arrival of R-L151 (specifically R-L21) into the British Isles is abrupt, traced in England by ancient DNA beginning 2383 BC (I5379), and seemingly reaching southern Scotland before 2163 BC (I2568) and Ireland before 1641 BC (Rathlin1). This transition corresponds with the arrival of the Bell Beaker groups and the arrival of metallurgy, with autosomal DNA identifying a $\gtrsim$90% population replacement beginning around 2450 BC[19].

With the exception of a single R-U152>L2 return (I2567), British and Irish ancient R-L151 remain uniquely R-L21 until the late Bronze Age, when R-DF27 individuals are found as well (from 1212 BC, I16454). These are presumed to come from the R-DF27 stronghold in Iberia via the Atlantic trade routes, but this period also corresponds to a (related?) rise in autosomal DNA linked to the Eastern European Farmer community[20].

The R-U106 fraction in these initial waves of migration is statistically zero. A single individual, I11149, is found in Cambridgeshire, recovered from a shallow grave with no grave goods in the vicinity of an Iron Age ditched enclosure dated 733–397 BC[21]. This individual is R-Z156, with possible calls for R-Z156>A9555>Z5889. The lack of other burials in pre-Roman Britain caps the R-U106 fraction across the UK and Ireland in the few centuries surrounding 150 BC at below ∼3% (95% probability) and in England specifically at below ∼3.5%. While we can be less precise in Scotland, we can constrain the R-U106 population of Scotland during Iron Age and early Roman times to <11% (68% confidence). Ireland and Wales do not have enough individuals from that period to determine a meaningful limit individually.

This absence of R-U106 picture changes considerably during the Roman occupation of Britain. There are only seven Roman-era burials with Y-DNA calls in our sample, from which it is impossible to obtain accurate statistics. However, two of the seven are U106+. 6DT3 and 3DT16 belong, respectively, to R-Z156>DF98>S1911>S1894>S4004>FGC14814 and R-Z156>DF96>S11515>L1[22]. These individuals were excavated in Driffield Terrace, York, from a late Roman-era (circa 300 AD) cemetery. Their strontium and oxygen isotope ratios are typical for the local region. Their burial (including decapitation) is consistent with deaths as gladiators, or otherwise a military rôle.

We can therefore summarise the Roman and pre-Roman phases of ancient DNA in Britain with the statements that R-U106 occupied an extremely small fraction of the population compared to today. All three burials are from R-Z156: this comprises only 18% of the R-U106 tests in the Family Tree DNA database, and 19% of British/Irish R-U106, suggesting that the small trickle of R-U106 individuals who did make it to England's shores before the fall of Rome were dominated by R-Z156 sub-clades. The vast majority of R-U106xZ156 haplogroups with UK or Irish ancestry today must therefore have arrived after the Roman invasion, and likely after the retreat of the Roman empire from Great Britain in 410 AD. The presence of R-Z156 groups during the Iron Age and Roman Empire is commensurate with their arrival during these periods from regions of north-west Europe that are Z156-rich.

In the post-Roman era, we see many ancient U106+ burials, with the fraction in England around 500 AD reaching a peak of 24–47% (95% c.i.). This compares with a modern percentage of 22 ± 9%[23], indicating that the post-Roman Germanic migrations were indeed the largest contributor of R-U106 in England, and possibly (via England) the wider British Isles.

Incoming haplogroups during this period are broadly representative of modern UK/Irish R-U106 fractions. Statistics of individual haplogroups are too small to identify any notable absences, and no R-U106 sub-clade is present in surprisingly large numbers among the sampled sites in East Anglia, Wessex, Deira or Kent. Insufficient Viking or Norman burials exist to determine the contribution of these groups to modern British R-U106.

5.5.3 Nordic countries

Ancient DNA in the Nordic countries oversamples Denmark and southern Sweden. Co-incidentally, these are the regions of the Nordic countries where R-U106 is strongest today.

R-U106 in the Nordic countries does not seem to begin simultaneously with the arrival of the Corded Ware Culture (in this region, the Single Grave Culture, circa 2850 BC[14]). Instead, they belong to the R1a component of the migration, R1a-PF6162 (RISE61, RISE94, ber1M, CGG107425, Oslund; 2672–2356 BC).

This abruptly changes around 2300 BC, when burials start to become dominated by either R-U106 burials, or untyped upstream burials from R-L11* and R-L51*. These begin in the Jutland peninsula from 2290 BC (NEO870), passing rapidly over the Danish Islands (NEO92, CGG106770; R-L11) to southern Sweden (RISE98; 2154 BC). By comparison, R-P312 (R-L21) is not seen in Scandinavia until ∼1150 BC (NEO946). This influx of R-U106 is concurrent with the arrival of the Bell Beaker Culture (circa 2300 BC[14]).

The content of these Bronze Age burials, all from modern Denmark and southern Sweden, is vastly dominated by R-Z18 (nine burials). A couple of other basal clades of R-U106 also make an appearance, from RISE98's unique

haplogroup that appears to gone extinct, to CGG106838 (R-Z301>FGC13959, pre-S9891; 2281–2048 BC). R-Z18 includes 37% of R-U106 testers in the Nordic countries today, and 14% of R-U106 testers in Europe as a whole. The probability of observing nine out of 11 R-U106 burials being R-Z18 with a 37% R-Z18 frequency is 0.04%, thus R-Z18 was a much larger fraction of R-U106 in the Nordic countries than it is today (probably about 73–93% of R-U106 [68% c.i.]).

The foundation of R-Z18 also lies close to 2300 BC (2608–1997 BC; 68% c/.i.; CGG107465 constrains it to before 2026 BC). It therefore seems reasonable to place the origin of R-Z18 in Jutland, coincident with the arrival of the Bell Beaker Culture into the region and the development of the Nordic Bronze Age. Conversely, the high R-Z18 fraction among R-U106 means we may place much of the remainder of R-U106 south of Denmark during this period.

The fraction of R-U106 in the region is still poorly constrained: the tested U106+ fraction reaches a maximum in Denmark around 1900 BC, representing 28–61% of samples (68% c.i.). This should be treated as a lower limit to the fraction of the Danish population who were U106+, due to the number of R-L11* and R-L51* burials, and the lack of R-L51xU106 burials. (The latter's absence increase the likelihood that R-L11* and R-L51* burials are R-U106 and, specifically from the above, R-Z18).

The composition of Nordic (especially Danish) DNA changes again around 300 BC, where a more pan-European set of R-U106 sub-clades arrives. While R-Z18 remains abundant, new arrivals include R-S18632, R-Z9 (R-Z7), R-FGC396, R-Z159 (R-CTS6353), R-S12025, R-U198. The arrival of new groups (R-L47, R-Z326, various under R-Z8) continues during the Viking period and later.

5.5.4 North-west Europe

Variations in the R-U106 fraction in north-west Europe over time can partly be attributed to uneven sampling. In particular, there a large number of French Celtic burials during the period 675–100 BC, which are not necessarily representative of north-west Europe overall. However, lack of numbers in any individual country preclude examination of the data at higher spatial resolution, except at a few key time steps.

The first R-L11 burial in north-west Europe is found in Switzerland (Aesch25, 2685 BC). Early burials in north-west Europe are vastly dominated by R-P312, especially R-U152 (beginning with RISE563 in south-east Bavaria, 2572–2512 BC). The first R-L21 is seen in south-west France by 2461–2299 BC (GBVPK).

The first R-U106 in north-west Europe is much later (1911–1766 BC), when a young adult from the Únětice culture (LEU007) is found in Thuringia, south of the Harz mountains in Germany.

R-U106 makes its first appearance in south Holland at some point in the range 2136–1892 BC. I13025 was a youth in the Barbed Wire Beaker culture, a northern offshoot of the Bell Beaker Culture. This burial is not typed beyond U106+. This is one of four Bronze Age burials in Holland, with the later three being I4070 (1880–1657 BC, R-Z301?Z304), I11972 (1501–1310 BC, R-Z381xZ301, therefore likely R-Z156) and I17019 (1421–1216 BC, R-Z381>Z156>Z304). This establishes a strong and even dominant (>50%) R-U106 presence among the local population of Holland by ∼1700 BC, and at least some R-U106 presence in the centuries before. The major R-U106 sub-clade in this group is likely to be R-Z156 and possibly R-Z304 in particular. R-Z156 continues to have a dominant presence in the Netherlands during the Roman period (CGG107754, CGG107735, CGG107751 [R-Z304>DF96>FGC13326>S25234]; also CGG107767 [R-Z381]).

R-U106 does not make an appearance in French ancient DNA until the Iron Age La Tène Culture (740–390 BC), when a single adult male (COL239, R-Z156>S3311) is found north-east of Paris. However, ancient DNA between 1400 and 700 BC is very lacking due to cremation burials predominating during this period, meaning absence of evidence is not evidence of absence. R-Z9 is also found in southern France during the La Tène period (CLR23).

This picture presents a slow migration into north-west Europe, probably spread westwards during the time of the Únětice culture, reaching central Germany and the Netherlands before around 2000 BC, and becoming a significant component of both the Tumulus and the Nordwestblock (Elp, Hilversum, etc.) cultures of north-west Europe. Later Celtic or pre-Celtic cultures could then take it further afield to France and England. The dominance of R-Z156 in these movements in clear, and probably represents the dominant component of R-U106 on the route spreading west and south-west from Bohemia.

5.5.5 North-central Europe

Bronze Age ancient DNA from north-central Europe is dominated by the Czech Republic, with a few samples from Poland. We can therefore make strong statements about R-U106 in the Czech Republic, but few statements about other countries in the region.

The aforementioned PNL001 represents the beginning of the Corded Ware Culture in eastern Bohemia. However, the R-U106 fraction in the region remains very low during the third millennium BC (2000 BC ± 500 years, Czech Republic: <8%, 95% c.i.). This indicates that, while R-U106 was clearly already present in the Czech Republic, it is not a dominant haplogroup. Therefore the majority of R-U106 probably did not stay long in Bohemia.

Several R-L11* results are found in Bohemia around 2770 BC (STD002, VLI092, VLI011), but the dominant group in the area appears to have been R1a-M417, mixed with some I2a-Z161, which presumably descended from the earlier Globular Amphorae Culture. R-U152 then arrives with the Bell Beaker Culture (first result I7278, circa 2383 BC), which then goes on to dominate during the Bell Beaker period.

The individual I7196 represents an important cornerstone in tracing the migration of R-U106 and establishing absolute dates on the haplotree. I7196 was an older (40+) individual, found in a suburb of modern Prague. His grave was one of several found at the site, with posture and grave goods consistent with the early part of the Únětice culture (circa 2200–1950 BC). His haplogroup has been demonstrated down to R-Z304, and he has single reads for S1911 and S1894, so is likely R-S1894, but officially he remains R-Z304?S1911?S1894. The age of these haplogroups (as derived by Family Tree DNA) is 2777–1621 BC for R-Z304, 2439–1284 BC for R-S1911 and 2262–1095 BC for R-S1894. We can therefore determine that the age of these haplogroups must be in the earlier part of this potential range.

The dominance of R1a, I2 and R-U152¿L2 continues throughout the Únětice period. The archaeological record is scant during the Tumulus Culture due to burial practices, but become present again around 1000 BC.

- In Bohemia: the Urnfield Culture (1300–800 BC; I13788; R-Z156>Z304) and La Tène Culture (480–390 BC; I15950; R-Z304>BY12480>BY12482/Y28944).
- In Slovenia: a Hallstadt culture infant (C/D period; 742–400 BC; I23978; R-Z156>S5520>FT221936).
- In Austria: an adult in Hallstatt (750–450 BC; CGG101214; R-U106) and a Batavi burial (26–126 AD; R10659; R-Z156>FGC39800>FGC39815>BY126375).

R-U106 represents a modest proportion of burials during the Iron Age (500 BC Czech Republic: 10–40%, 68% c.i.), and probably higher percentage than during the Bronze Age. The prevalence of R-Z156 even into the Roman period is stark, emphasising the trend in north-west Europe and England (see above) that the southern margin of the R-U106 distribution is predominantly R-Z156 until the Migration Period, when R-L48 groups become most common.

5.5.6 Eastern Europe

Ancient DNA is eastern Europe has very little R-L11 content. An Iron Age Scythian in modern Ukraine has been tested R-L2 (scy009), which represents the only R-L11 burial until the Viking Age. The majority of ancient DNA testing has been done in Russia. The Baltic States have been covered proportionally to their size and population, with Estonia being well tested in the middle Bronze and Viking ages; the Ukraine remains under-tested, while Belarus is absent from ancient DNA studies.

5.5.7 South-east Europe

South-east Europe is moderately well tested for ancient DNA, with particular emphasis having been put on testing in Hungary. R-L11 is largely absent, except for a few sporadic burials starting in the early Bronze Age (I2365, R-L2 and I2365, R-L11). R-L2 remains common during the later Bronze Age, among haplogroups E1b, J2 and G2a. The first R-U106 are not seen until the Migration Age (Langobards SZ2, SZ11 and SZ4).

5.5.8 Mediterranean

The Mediterranean has very low rates of R-U106, though R-P312 is clearly present as expected. Two R-U106 Bell Beaker individuals have been found in Spain[24] but were published after the sample used here as a basis. CGG_2_023808 dates from the early phase of the Bell Beaker group (2115–1762 BC) and is Y3444+ (found in R-FGC396, but is possibly better described as pre-FGC396). CGG_2_023745 dates from the later phase (1619–1462 BC) and is R-S18632. Both of these haplogroups are minor clades of R-U106, showing that some R-U106 did become entrenched in the Bell Beaker culture, but did not thrive in it in the same way that the major R-P312 clades did. Modern populations of R-FGC396 and R-S18632 do not concentrate in Spain, so it is unlikely that Spain represents a point of origin for either R-FGC396 or R-S18632.

5.5.9 Conclusion

From this data, we have a coherent picture of R-U106 spreading from in or near Bohemia in the early phases of the Corded Ware Culture (approximate period 3050–2900 BC). R-Z18 is then found approximately 600 years later in modern-day Denmark. Meanwhile, R-Z156 expands slowly across Europe, to be bound by the North Sea and the Alps, until it first crosses into England in significant number at some point during the first millennium BC. Earlier migrations by minor clades (R-FGC396, R-S18632) seem to have occurred.

R-L48 and R-S1688, although representing half of R-U106 today, remains conspicuous by its absence in the ancient DNA record until a single burial in the La Tène culture. It then shows up in Denmark 2000 years ago, before becoming widespread during the Migration Period. The absence of R-L48 and R-S1688 in ancient DNA cannot be solely down to their being smaller in the past, so we should examine the possibility that R-L48 (and possibly the smaller R-S1688) predominantly existed in cultures practicing cremation burials from which ancient DNA cannot currently be recovered (e.g., the Urnfield Culture) and regions where soil conditions do not preserve ancient DNA. They would then "break out" of these regions during the Migration Age.

We can therefore expect R-L48 (and maybe R-S1688) to have existed in the regions with gaps or large uncertainties in Figure 3. Nevertheless, we know that immediately before the Migration Age, many R-L48 groups are likely to have been in the regions from which the Germanic Tribes took over the fallen Roman Empire, since it was from these

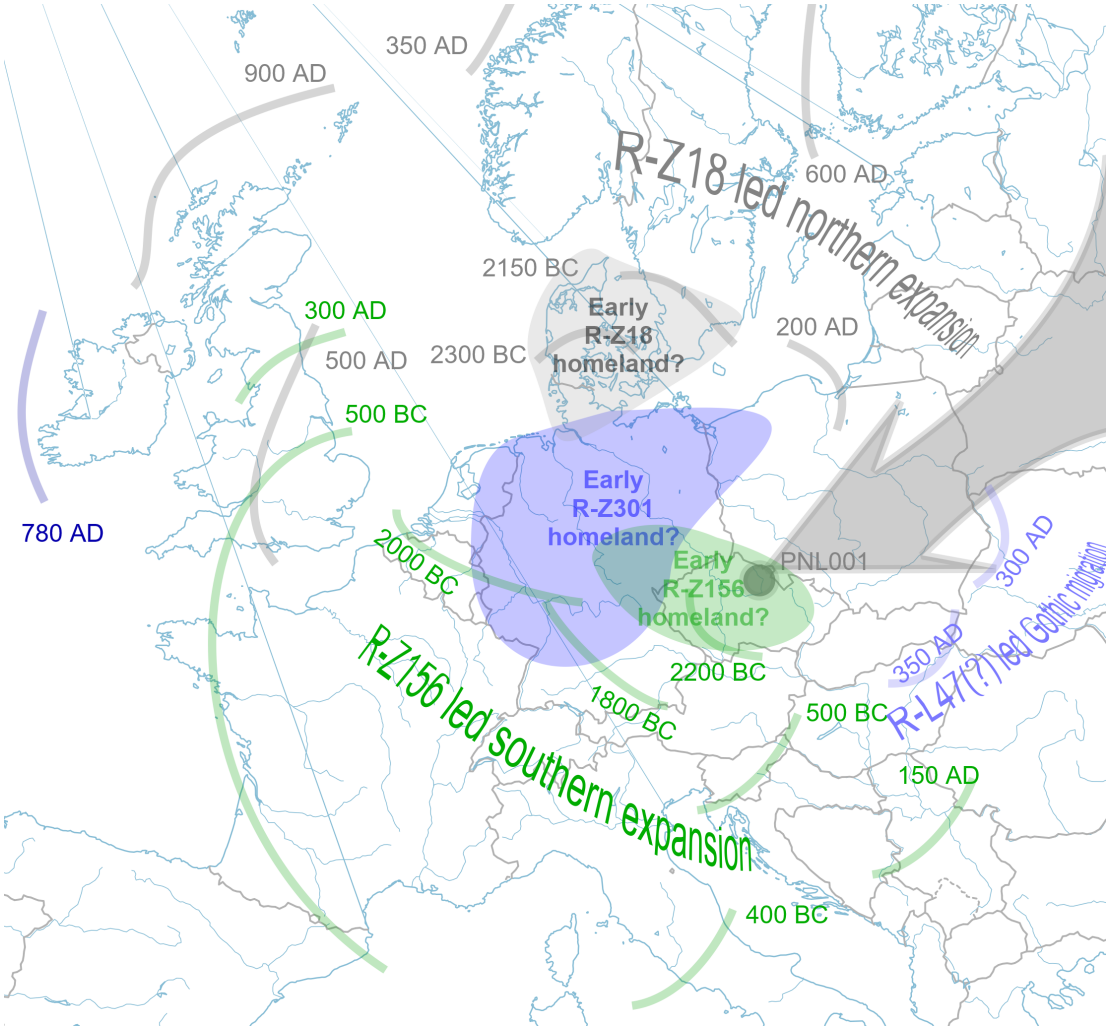

Figure 4: Generalised model of the *latest* possible dates of R-U106 expansion and regions where early R-U106 groups might have lived *as derived from ancient DNA alone*. Arriving with the Corded Ware Culture and presumably establishing itself near the burial PNL001, a southern front radiates outwards containing mostly R-Z156, while a northern front simultaneously expands into the Baltic containing mostly R-Z18. R-L48 and R-S1688 are largely absent until the Iron Age, hypothesised to be residing in the under-tested parts of Germany, during which time a notable expansion (containing at least some R-L47) progresses in the east.

positions that we source the R-L48 ancient DNA found during the Migration Age. While recognising the dangers of assigning an origin to the most-populous or most-diverse region, it is also likely that R-L48 and R-S1688 arose in a region where one or both haplogroups are still commonly found today. The obvious candidate that fulfils all of these criteria is to place the R-Z301 ancestor (which ties R-L48 and R-S1688 together) in modern-day northern and/or western Germany. This summary is illustrated in Figure 4.

6 Phylogenetic methodology

To understand the phylogenetic spread of the R-U106 family, we therefore need to step beyond ancient DNA, using its results in combination with information from modern testers, while avoiding the pitfalls mentioned in Section 3.

Any meaningful methodology must make use of the tree-based structure for at least the upper portions of the tree. However, conventional tree-based structures fail for the smallest groups since, when numbers become statistically small, conventional structures only take into account the locations where positive tests are found, without considering locations where untested populations could be. This is a problem particularly for heavily biased sampling, as it tends to force an origin for small haplogroups onto well-tested locations, which can cause subsequent issues higher up the tree when these smaller haplogroups combine to form larger ones.

In the following, a manual approach is taken. Each sub-clade of R-U106 is processed, smallest to largest, from the root of the tree to the present day. Country-level counts are generated for the entire haplogroup and any major sub-clades are computed. Relative frequency statistics can be generated from the country-level data, to determine how the haplogroup is distributed compared to its parent, its component sub-clades, R-U106 as a whole (and any other haplogroup of interest). From these statistics, it can be assessed which sub-clades are individually big enough to measure (see final comments in Section 2.3). For those large enough, absolute statistics can be corrected for bias and a median position for the haplogroup computed. The following questions are then posed.

1. Are there any obvious indicators of origin for this haplogroup (e.g., ancient DNA from close to the haplogroup's TMRCA, known historical genealogies)?
2. Are there any obvious indicators of origin that can be used from nearby haplogroups (e.g., since both PNL001 in R-U106 and I7196 in R-Z304?S1911?S1894 are in Bohemia, it's more likely that lineage stayed in Bohemia during the intervening centuries).
3. If there is more than one major sub-clade, do the sub-clades have the same geographical distribution? If so, there was probably no major migration around that haplogroup's formation (though see issues with relative timing of migrations in the example of R-L151). If not, a migration near to this node in the haplotree is likely.
4. Are there any major founder effects in the sub-clades that need taken care of before statistics are drawn up (including heavily tested individual families)?
5. How likely is it that untested populations could be missing from the haplogroup, and how could untested populations affect efforts to ascribe an origin?
6. Are the individual statements on earliest-known ancestor information greatly different from the distribution observed in country flags? Do they provide any information on levels finer than a country-level scale (e.g., are returns common in a particular place within a country)?
7. If a haplogroup is smaller than its contemporaries, does it share a similar geographic distribution with any (e.g., for small haplgroups in around 2000 BC, do they geographically appear today to be R-Z18-like, R-Z156-like or R-Z301-like)?
8. If major sub-clades do not share a distribution, do they overlap? If so, the origin is more likely to be near the overlapping region.
9. Do the minor sub-clades exhibit a different geography to one or more major sub-clades? If so, the haplogroup might have had a successful migration, leading to the major sub-clades, while the minor sub-clades better represent the origin (cf., R-L151 versus the smaller R-M269xL151 sub-clades).
10. Are there any better constraints than Family Tree DNA's estimates on the TMRCA? If so, a new TMRCA can be generated.
11. Can the TMRCA and any proposed location be linked to a known historical or archaeological culture/migration?

Statistics we have to address these questions include:

1. Individual country and region counts, which can be compared between haplogroups.

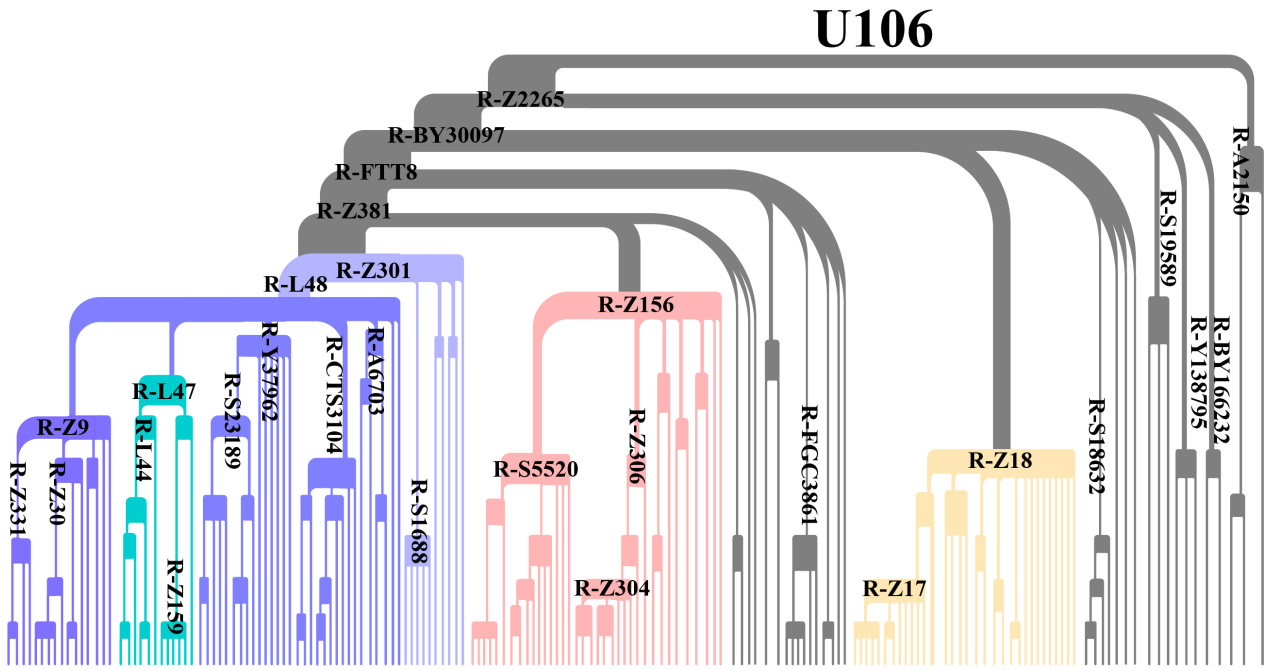

Figure 5: Upper tree of R-U106, showing the first 15 SNPs worth of expansion. This covers the approximate period 3000–2000 BC.

2. Bias-corrected mean locations, computed as:

$$< (\text{Lat}, \text{Long}) >= \sum \frac{Nb(\text{Lat}, \text{Long})}{Nb} \tag{2}$$

where $(\text{Lat}, \text{Long})$ are the midpoints of each country, N the count in that country, and b the bias factor of each country. These show the modern mean locations of individuals. Notable locations for major haplogroups are:

- R-Z2265: 50.37 N, 10.32 E; central Germany.
- R-Z156: 49.30 N, 7.58 E; Franco–German border.
- R-Z18: 51.06 N, 12.63 E; near Leipzig, eastern Germany.
- R-Z301: 50.60 N, 10.77 E; Thuringia, central Germany.
- R-P312: 47.70 N, 4.93 E; near Dijon, central France.
- R-L21: 50.14 N, 0.89 E; English Channel.
- R-U152: 47.24 N, 10.28 E; westernmost Austria.
- R-DF27: 45.82 N, 2.98 E; near Clemont-Ferrand, south-central France.

These mean values can be improved using individual latitude and longitude combinations where available.

3. Intra-country median locations, computed similarly, which are mainly useful for correlating with historical migrations in large and well-populated countries/regions like the British Isles.

4. The two-dimensional Kolmogrov–Smirnov test[o] for individual latitudes and longitudes, which can be used to find differences between haplogroups (strictly, the probability that both haplogroups are not drawn from *exactly* the same population). Low P values do not necessarily mean different origins, since this test does not take into account differential migration after the haplogroup formed.

7 Detailed analysis of the origin and spread of R-U106 and its sub-clades

7.1 Minor near-basal clades of R-U106: early expansion

7.1.1 Context

Due to the exceptionally early nature of PNL001 in the R-U106 and Corded Ware Culture chronology, we have essentially adopted the location of PNL001's burial as the original location from which R-U106 spread, and the immediate forebears of PNL001 as the founders of R-U106 (Section 5). This is likely a simplification of the true

[o] `https://github.com/syrte/ndtest`

nature of R-U106's expansion, but should be sufficiently close that we can use it as a meaningful starting point for interpreting haplogroups further down the R-U106 tree. An assumed TMRCA for R-U106, based on a coverage of 23 Mbp and one SNP since R-L151, is 3020 BC (68% c.i., 3146–2906 BC; 95% c.i., 3288–2777 BC).

The following analysis is based on the 2024-Dec-23 version of the Family Tree DNA haplotree. The upper regions of this tree can be seen in Figure 5. The major haplogroups R-Z18, R-Z156 and R-Z301 do not appear until a considerable way down this tree, before which it is dominated by what are now minor clades within R-U106. R-Z301 takes up 36% of branches (modern size: 64% of testers), R-Z156 occupies 27% of branches (modern size: 19% of testers) and R-Z18 occupies 17% of branches (modern size: 13% of testers). Minor near-basal clades, shown in grey, occupy the remaining 20% of branches (modern size: 4% of testers). We can see from these percentages how the minor clades have waned in size while R-Z301 in particular has grown.

The R-Z18 growth appears a distinct event, discontinuous to the growth of other haplogroups around it. This suggests that R-Z18 followed a markedly separate path from the start, and almost died out.

The growth of R-Z156 shows an early dominance by R-S5520. This is now dwarfed by both R-DF96 and R-DF98, which only form at the very bottom of this tree, under R-Z304.

In R-Z301, the sizeable R-U198 sub-clade represents a much later growth to the tree than shown here. Similarly large haplogroups like R-Z8 and R-Z326 have yet to make an appearance. This leaves R-L48 dominated by now relatively minor haplogroups like R-Y37962 and R-CTS3104.

This tree therefore gives a rough expectation of the haplogroups to be found in ancient DNA between 3000 and 2000 BC. This explains why minor clades are found more commonly in ancient DNA, but emphasises the surprising lack of R-Z301 burials, which should represent ∼36% of the R-U106 burials in the latter part of the third millennium BC.

7.1.2 R-U106>A2150

TMRCA: Assuming two SNPs since R-U106 at a coverage of 14.3 Mbp, 2780 BC (95% c.i., 3093–2380 BC).

Ancient DNA: LEU007 (1911–1766 BC), Únetiče culture, Thuringia. R-A2150>pre-BY69794): U106+ (2 reads), FT420436+ (1 read), ZS1682+ (1 read), BY69794- (3 reads), BY74465- (3 reads).

Modern testers: There are 11 testers with known European origins, nine of whom are from the British Isles and two from Germany. This is too few to derive meaningful statistics. No further useful indications from project matches.

Expansion: The main periods of expansion are the initial growth of the group around 3000 BC, and during the early medieval period.

Conclusion: LEU007 indicates that at least some early R-A2150 belonged to the western extremities of the Únetiče culture in what is now modern Germany. If the haplogroup stayed roughly in this region, that would be sufficient to explain the modern testers in Germany, and its expansion during the post-Roman Germanic migrations.

7.1.3 R-U106>Z2265

TMRCA: Assuming two SNPs since R-L151 at a coverage of 23 Mbp, 2966 BC (95% c.i., 3215–2717 BC).

Narrative: Primary branch of R-U106, containing 99.96% of testers haplotyped below R-U106. Cannot be separated from R-U106 as a whole. Z2265 is expected to have occurred on the Corded Ware Culture migration westward from the R-U106 homeland.

7.1.4 R-U106>Z2265>BY166232

TMRCA: Assuming ten SNPs since R-L151 at a coverage of 14.3 Mbp, 2043 BC (95% c.i., 2463–1417 BC).

Modern testers: There are only three testers with known European origins, two French and one Slovakian. The Slovakian tester forms his own basal clade; the French testers are related in ∼1250 AD.

Expansion: The only major split is in 350 BC.

Conclusion: No significant information on origins.

7.1.5 R-U106>Z2265>Y138795

TMRCA: Assuming ten SNPs since R-L151 at a coverage of 14.3 Mbp, 2134 BC (95% c.i., 2575–1593 BC).

Modern testers: There are 12 testers with known European origins, only one of whom are from the British Isles. Of the remaining, there are three from Germany, one from Poland, two related testers from the Czech Republic, three related testers from Sweden and two unrelated testers from Spain. This is too few to derive meaningful statistics. The Spanish testers are basal, and could either represent early migrations to Spain, or later (probably post-Roman) migrations. No further useful information from Y-STR matches.

Expansion: An initial period of expansion around 2000 BC is followed by a second expansion between 250 BC and 400 AD.

Conclusion: The unclear origins of the Spanish testers cloud judgement of the origins of R-Y138795. Lack of British testers suggests avoidance of western European areas (sources of Angles, Saxons, Belgae, etc.). Its locus appears to be in central Europe, with or without an early migration to Spain.

7.1.6 R-U106>Z2265>S19589

TMRCA: Assuming five SNPs since R-U106 at a coverage of 16 Mbp, 2486 BC (95% c.i., 2839–2038 BC).

Ancient DNA: HID004, 421–537 AD, Merovingian Frank from modern Hanover. R-S19589>FGC57430>BY116631.

Modern testers: 28 European testers, 15 from the British Isles. The remainder are from France, Germany (3 with 2 historically related), Poland (2), Denmark, Sweden (4), Russia and Lithuania. The distribution shows marginally better affinity with R-L48 than R-Z18, suggesting a more central European locus, but with low confidence. Its presence in Russia appears to be historical immigration of German families.

Expansion: Some expansion during the period 1000–600 BC.

Conclusion: This haplogroup is well spread across the northern bounds of R-U106's overall distribution and covers the entire gamut of Germanic countries. A suggested locus during the Iron Age is among the early Germanic people of the Jastorf culture or (given the timing of its expansion) its immediate predecessors.

7.1.7 R-U106>Z2265>BY30097

TMRCA: Assuming three SNPs since R-L151 at a coverage of 23 Mbp, 2916 BC (95% c.i., 3173–2645 BC).

Modern testers: One basal tester from Italy.

Narrative: Primary branch of R-U106>Z2265, containing 99.72% of testers haplotyped below R-U106. Cannot be separated from R-U106 as a whole. BY30097 is expected to have occurred on the Corded Ware Culture migration westward from the R-U106 homeland.

7.1.8 R-U106>Z2265>BY30097>A10122

TMRCA: Family Tree DNA provides 1318 BC – 98 AD (95% c.i.). No significant improvement is realistic from ancient DNA.

Modern testers: Nine testers, five with European origins, all in the British Isles. Three of the five are from the Gillespie family of southern Scotland. The other two are from south-west England, related 2000 years ago.

Narrative: While not impossible, it is unlikely that the origin of R-A10122 is in the British Isles, yet this is the only place for which we have data on this very small haplogroup. No further information regarding origins is possible.

7.1.9 R-U106>Z2265>BY30097>S18632

TMRCA: Assuming 11 SNPs since R-BY30097 at a coverage of 20 Mbp, 2256 BC (95% c.i., 2615–1841 BC).

Ancient DNA:

- CGG_2_023745, 1619–1462 BC, Bell Beaker, near Granada, southern Spain.
- CGG023274, 389–206 BC, Iron Age Jutland (Denmark).
- CGG107498, 1–200 AD, Iron Age Zealand (Denmark).

Modern testers: 197 modern testers, 65 with known European origins. Of these, 36 are from the British Isles. Of these, 18 are from England, which is typical of the ∼50% found in other R-U106 groups. Within these, there is the large English R-BY61415 Bell family (26 of 197, 7 of 65, 6 of 36) and the large Northern Irish R-FTC19087 Paisley family (15, 5, 5), which have the potential to distort the statistics of the haplogroup overall.

R-S18632 splits into a young (950 BC) branch, R-S11320, and an older branch, R-Y15798, that formed part of the initial R-S18632 expansion. Both of these sub-clades show similar distributions within Europe: both show up in Germany (2 and 4 testers, respectively), the Netherlands (2,3) and north-central Europe (Poland: 1,2; Czechia 0,1). R-Y15798 also has a component in Denmark (3) and Sweden (6), Latvia (1) and the Ukraine (2). A Hungarian tester is also undifferentiated within the R-S11320 group, with a Y-STR match in modern Poland.

The mean location is typical of R-U106 haplogroups (Kassel, central Germany), but the population does not show significant affinity with R-Z18, R-Z156 or R-L48 as a whole. Two of the Dutch testers are closely related. The Scandinavian testers belong to four sub-clades within the haplogroup, each related internally around 200 BC or later. The north-central/eastern Europeans occupy four distinct groups, each related internally around 500 AD or later.

Expansion: An initial period of expansion lasts for a short time, but fails to grow a large haplogroup quickly. A period of contraction likely occurs until 1200 BC, when slow, continual growth starts. Peaks in growth occur around 1200 BC, 600 BC and from 400 AD.

Conclusion: CGG_2_023745 indicates some participation in the Bell Beaker groups of which R-P312>DF27 was likely a part. The most likely point of contact for these groups is probably the Rhineland Beaker groups, so an initial spread west is suggested. The likely TMRCA range covers this potential encounter nicely.

Migrations to Scandinavia and eastern Europe show clear groupings that may indicate later migrations. The Germans and Dutch in the group do not show such clear groupings, so it is suggested that this region represents

a better origin for the haplogroup (Nordwestblock cultures?), with Iron Age expansion into Scandinavia (Jastorf culture?) and later migration (early medieval?) into eastern Europe.

Notable subgroups:

- R-S18632>BY65802: appears to be strongly Scandinavian. Their common ancestor lived around 800 BC, but branching around 300 BC may indicate that this is the point of migration into southern Scandinavia or growth from a very small population (again in southern Scandinavia). Speculatively, that migration/growth could be as part of the rise of the Germanic peoples, extending out of the Jastorf culture into modern Denmark and southern Sweden.

7.1.10 R-U106>Z2265>BY30097>S12025

TMRCA: Assuming 14 SNPs since R-BY30097 at a coverage of 20 Mbp, 1946 BC (95% c.i., 2340–1504 BC).

Ancient DNA: CGG106796, 1–400 AD, Iron Age Zealand (Denmark).

Notes: Separates cleanly into the younger R-S16361, which makes up slightly less than half of the haplogroup and the older R-FGC12021, which formed during the initial expansion. These are dealt with separately.

Conclusion: Both the R-S16361 and R-S25007 sub-clades show a split between the west and north Germanic groups. While this could be co-incidence, it could also represent a migration of part of the R-S12025 group some time after its foundation that encompassed both groups, leaving populations in both. Analysis of the sub-clades below suggests that this migration was over 1600 years ago, so probably dates to at least the early Germanic period, if not before. The presence of this haplogroup in modern Denmark only after the Iron Age could link its migratory history with that of R-L48.

7.1.11 R-U106>Z2265>BY30097>S12025>S16361

TMRCA: Family Tree DNA provides 408 BC (95% c.i., 931 BC – 27 AD). Ancient DNA provides no further constraint.

Ancient DNA: GRO008, R-S16361>S19367, Groeningen (NL).

Modern testers: 85 testers, 56 with a known European origin, 32 in the British Isles. This includes the large R-FGC15048 Gordon–Seaton family (27/85, 20/56, 20/32), which bias the statistics. The continental population is strongly Dutch (10), with some Germans (3). It is also strong in Scandinavia (6 in Sweden, 1 in Denmark, Norway, Finland). Individuals in Russia and Portugal are also known. The Dutch R-FT248930 is ∼1500 years old, but the wider R-BY165382 appears mostly Scandinavian.

Expansion: The main period of expansion appears to be between 200 AD and 500 AD or shortly thereafter.

Conclusion: This haplogroup has a strong and persistent connection to the Netherlands for at least the last 1500 years. However, it also has had strong connections to Sweden and the wider parts of southern Scandinavia (and the Germanic world) that go back at least as far. Its origin probably lies in one of these places. This haplogroup is therefore very likely dominated by the Germanic peoples, though we cannot definitively claim it was founded among them.

7.1.12 R-U106>Z2265>BY30097>S12025>FGC12021

TMRCA: Assuming 16 SNPs since R-BY30097 at a coverage of 20 Mbp, 1799 BC (95% c.i., 2210–1345 BC).

Notes: 100 testers, 51 with a known European origin, 42 in the British Isles. The small, basal R-FT205640 branch is traced by a modern German family. A number of near-basal clades spawn Irish (and some Great British) families, though testing biases and consequent incompletenesses mean these cannot be taken as these haplogroups' point of origin. The origins of these (near-)basal clades are unclear, but this reduction leaves R-S25007 as a significant point of European origin.

7.1.13 R-U106>Z2265>BY30097>S12025>FGC12021>S25007

TMRCA: Family Tree DNA gives 746 BC (95% c.i., 1275–300 BC).

Ancient DNA: HAD009, R-S25007>FGC31905>FGC53757. A young Angle male buried in Cambridge during the 5th/6th Century AD.

Modern testers: 83 testers, 41 with a known European origin, 34 in the British Isles. The Fuller family (emigrants on the *Mayflower*) make up 38/83, 18/41 and 18/34 of these.

Expansion: There are three basal clades. The Dutch-dominated R-S11595 expands from circa 1000 AD. The Scandinavian-dominated R-BY49031 expands from circa 400 AD. R-FGC31905 represents part of the initial expansion and retains only one German tester and HAD009 for guidance, but shows significant expansion between 100 BC and 300 AD.

Conclusion: This haplogroup's distribution appears Germanic, but it is again not clear whether the Netherlands (west Germanic) or southern Scandinavia (north Germanic) is a likely origin. HAD009 suggests a presence in the Angle homeland by 400 AD.

R-U106>Z2265>BY30097>Z18

This haplogroup is dealt with in its own section.

7.1.14 R-U106>Z2265>BY30097>FTT8

TMRCA: Assuming four SNPs since R-L151 at a coverage of 23 Mbp, 2868 BC (95% c.i., 3133–2579 BC).

Narrative: Primary branch of R-U106>BY30097, containing 86.2% of testers haplotyped below R-U106. Cannot be separated from R-U106 as a whole. FTT8 is expected to have occurred on the Corded Ware Culture migration westward from the R-U106 homeland. The locus of R-FTT8 is moved south to counteract the comparatively northerly movement of R-Z18.

7.1.15 R-U106>Z2265>BY30097>FTT8>FT421644

This haplogroup has sole representatives among a modern Spanish family. This family could either be truly anciently Spanish (since the origin of the haplogroup), or they could represent a more recent immigration.

7.1.16 R-U106>Z2265>BY30097>FTT8>FT44298

TMRCA: Assuming five SNPs since R-FTT8 at a coverage of 14.3 Mbp, 2386 BC (95% c.i., 2783–1860 BC).

Narrative: A single tester of unknown origin diverges in an early branch. Two other testers, one of whom is German, are related via a medieval ancestor. Insufficient data to assert an origin, but highest probability is naturally in modern Germany.

7.1.17 R-U106>Z2265>BY30097>FTT8>FGC396

TMRCA: Assuming 12 SNPs since R-FTT8 at a coverage of 20 Mbp, 2113 BC (95% c.i., 2505–1658 BC).

Ancient DNA:

- CGG_2_023808, 2115–1762 BC, Bell Beaker culture, south-eastern Spain. Y3444+, no information on other calls in R-FGC396. Could be pre-FGC396.
- CGG019205, 2 BC – 54 AD, Iron Age Jutland (Denmark).

Modern testers / expansion: 83 modern testers, 33 European testers, 12 from the British Isles. The haplogroup's expansion was slow and it is dominated by the younger R-FGC403. The only R-FGC396xFGC403 tester is a R-FTC36275 Dutchman. The family of the US president Martin van Buren (1782–1862) is also R-FTC36275; Buuren is a town in Gelderland, suggesting R-FGC36275 has been in the Netherlands for probably the last 1500 years.

Conclusions: Establishing an origin for R-FGC396 is difficult without more basal testers. Much of its history must be established relative to the sub-clade R-FGC403. Its presence in Spain is not matched by modern testers, suggesting that CGG_2_023808 was from a failed R-U106 component of the Bell Beaker migrations to Spain.

CGG019205 shows that there was at least some Germanic component to R-FGC396, but this is not reflected in the R-FGC403 population, so CGG019205 may also represented a failed branch of the family.

Putting this together with the surviving R-FGC403 and R-FGC396xFGC403 populations, it seems plausible that R-FGC396 initally followed much the same course as R-S18632, entrenching itself into the Bell Beaker Culture and following into Spain and embedding itself into the cultures of north-western Europe during the late Bronze and Iron Ages. The R-FGC403 founder effect then skewed the mean location of this haplogroup towards the southerly end of its distribution.

7.1.18 R-U106>Z2265>BY30097>FTT8>FGC396>FGC403

TMRCA: Assuming 18 SNPs since R-FTT8 at a coverage of 20 Mbp, 1552 BC (95% c.i., 1997–1063 BC).

Modern testers: 63 modern testers, 26 European testers, eight from the British Isles. The R-BY153779 Bolen family comprise 10/63, 3/26 and 3/8 of these. R-FGC403 splits into the larger R-Z27230 and the smaller R-FGC415.

Both R-Z27230 and R-FGC415 show a strong absence of testers from the British Isles (8/26 = 31%, compared to the R-U106 average of 56%). There is also a probably absence of Scandinavians (2/26 = 8%, cf. 13% for R-U106), and these are a pair of closely related Finns. This leaves a dominant group in north-west Europe, with sporadic calls in the Czech Republic, Portugal and possibly Poland (the Poles are not tested below R-FGC396). Where detailed latitudes/longitudes are available, there is some possible concentration towards the Rhine.

This puts the bias-corrected mean location of modern R-FGC403 testers far to the south of most other R-U106 haplogroups, near Metz in eastern France. However, the distribution of latitudes and longitudes shows greater affinity for R-L48 (2D K–S test $P = 0.135$) than the southerly R-Z156 ($P = 0.0476$).

Expansion: A slow expansion persists until about 1000 BC, then the haplogroup remains dormant until an expansion period between about 100 BC and about 300 AD. The haplogroup then grows again from about 600 AD.

Conclusion: The lack of British and Scandinavian testers shows that R-FGC403 remained more localised than other southerly groups like R-Z156, presumably partly because it remained a smaller haplogroup for a longer time. The initial growth could be aligned to the Tumulus or Urnfield cultures, whose spheres of influence would fit the modern distribution.

Timings are approximate, so it is hard to ascertain hard facts from growth periods. However, taken at face value, no growth is seen during the main Celtic periods (700 – 100 BC). Growth is seen during the Roman period, but not the immediate post-Roman period. This suggests that the haplogroup benefited from the Roman Empire, either directly by being within its borders, or indirectly by taking advantage of neighbours weakened by the Romans. The lack of expansion during the Iron Age and post-Roman period suggests that the haplogroups did not take part in the Germanic expansions, so few of its members were part of the Germanic peoples. It is nonetheless difficult to place the main cohort of R-FGC403 during the Roman period within the bounds of the Roman empire, if Roman-era and later migrations to Great Britain are to be mostly avoided. A more Alpine location may be more favourable as the main locus of R-FGC403 during this period.

7.1.19 R-U106>Z2265>BY30097>FTT8>BY11501

TMRCA: Assuming 14 SNPs since R-FTT8 at a coverage of 20 Mbp, 2047 BC (95% c.i., 2434–1606 BC).

Modern testers: 93 modern testers, 49 with European origins, 31 from the British Isles. Vastly dominated by R-BY11507>BY11506, examined separately below. The R-BY11501xBY11506 testers are include a Swede and a medieval Czech family (Habarta), related to each other through the 1800-year-old R-BY68267.

Expansion: Slow to start.

Conclusion: The origin of R-BY11501 is difficult to determine, but an origin in modern Germany is consistent with the downstream R-BY11506 and may apply to R-BY11501 more broadly too.

7.1.20 R-U106>Z2265>BY30097>FTT8>BY11501>BY11506

TMRCA: Family Tree DNA provides 1345 BC (95% c.i., 1776–884 BC).

Ancient DNA: SED005, 650–875 AD, Norfolk. R-BY11506>BY50725.

Modern testers: 82 modern testers, 42 with European origins, 31 from the British Isles. The British Isles contingent contains R-BY72676, a medieval family from south-west England (historical Wessex), which is over-sampled, comprising 23/82, 6/42 and 6/31 testers. European testers derive from Germany (6), Denmark (2) and Sweden (3), though one of the Germans can only trace to US ancestry. The bias-corrected mean position of R-BY11507xBY72676 is in the Netherlands.

Expansion: Growth is fairly slow and continuous, though a slight peak occurs around 600 AD. The R-FTC13562 branch also shows rapid division after circa 1000 AD.

Conclusion: The slight growth during the post-Roman period, combined with ancient DNA from Norfolk, implies a component of this haplogroup in Saxon lands, though the ancient DNA connection is too old to apply this to any one branch of the haplogroup specifically. The Germans are well-dispersed throughout the haplogroup, while the Danish and Swedes are clustered into family groups, more firmly establishing Germany as this haplogroup's historic stronghold, suggesting a start in the Tumulus or Urnfield cultures.

R-FTC13562>Y30507 appears to be from the British Isles within the last 1000 years, while the basal R-FTC13562* tester is Swedish. This is suggestive of a Viking component.

7.1.21 R-U106>Z2265>BY30097>FTT8>FGC3861

TMRCA: Assuming 10 SNPs since R-FTT8 at a coverage of 20 Mbp, 2204 BC (95% c.i., 2563–1793 BC).

Ancient DNA: All ancient DNA is typed under the sub-clade R-Z8053.

Modern testers: 1132 testers, 288 with stated European origins, 196 in the British Isles. The three basal clades of R-FGC3861 (R-FGC14877, R-Z8053 and R-A1243) show significantly different distributions, which is partly down to enthusiastic testing by a few families and recent founder effects.

These numbers also include a large number of Family Finder testers who have not tested below R-FGC3861. The testers typed below R-FGC3861 represent only 475/1132, 170/288 and 125/196. Consequently, the country-level statistics for R-FGC3861 are listed here before consideration of individual subclades. The bias-corrected mean location (51.088 N, 14.557 E) is most comparable to R-Z18, but slightly further east.

Within the British Isles, 56% (109/196) are English, compared to 50% for R-U106 overall. While statistically significant (3σ), this is partly due to founder effects and testing of individual families. The fraction of Irish (R.o.I.+N.I.) testers (19/196 = 10%) is below the R-U106 average (16%).

In north-west Europe, there are 47 testers: 35 in Germany, 10 in France, 1 in Switzerland, 1 in the Netherlands. At least 11 of the Germans are from the R-FGC14877>BY39117 Ruth family, and one belongs to the R-A1243>BY200368 Pettit family of Suffolk.

In the Nordic countries, there are 16 Swedish, eight Danish, five Norwegian and two Finnish testers. Where recorded, these all belong to R-Z8053 (4/16, 2/8, 2/5, 0/2). Of these, all but one Dane are in the sub-clade R-S1855.

In north-central Europe, three Polish and one Austrian tester are recorded. At least one of the Poles are in R-A1243. Further east, four Russians are listed, again with at least one within R-A1243. In the south-east, one Hungarian and one Serbian are listed. In the Meditteranean, three Italians and one Portuguese are listed, with at least two Italians in R-Z8053.

Expansion: A sharp peak in expansion occurs around 700 BC. This occurs in all three sub-clades, but corresponds to the initial period of expansion of R-FGC14877. A broader peak occurs between 200 BC and 600 AD, with growth around 200 AD in particular. Considerable growth is again seen that peaks after 1000 AD, some of which is associated with the Norman de Verdun family.

Conclusion: The initial expansion of R-FGC3861 appears roughly in line with other R-U106 groups, but the disparity in the downstream haplogroups means that we need to look to their intersection to estimate an origin. This is made harder by the relative youth of two of these three groups, meaning we rely mostly on the basal clades of R-Z8053 to provide an origin location. This location has been placed in modern Germany, but is very uncertain.

7.1.22 R-U106>Z2265>BY30097>FTT8>FGC3861>FGC14877

TMRCA: Family Tree DNA provides 651 BC (95% c.i., 1224–174 BC).

Modern testers: 277, 70 with European ancestry, 51 of which are from the British Isles.

R-FGC14877 contains the R-BY39117 German Ruth family of the Rhineland–Palatinate[p] (11 German members), the R-A561 Booth–Allen families (3 UK, 4 English, 1 Irish members), and the R-FTD91363 Tryon family (12 members, 4 English, 1 French).

Given this information, there appears a slight excess of Scots in this group. Several of these Scots cannot trace their ancestry back to the UK. Those that can generally appear to be from the Borders or Central Belt. The Scots tend to group in clusters of no more than 500 years old. The English mean location (duplicates removed) is 52.65 N, 2.34 W (NW of Birmingham), with results stretching across western England (cf., the R-U106 English median of 52.16 N, 1.43 W, near Stratford-upon-Avon).

The other continental individuals are two French testers, a basal Danish tester and six other Germans. These other Germans are preferentially from northern Germany. One lists his latitude/longitude in Poland. A Portuguese and an Italian tester are also inferred from their Y-STR results.

Conclusion: R-FGC14877 splits fairly cleanly into German-dominated basal clades and English/Scottish-dominated R-FGC21340. A basal R-FGC21340>A563 German tester also exist. This suggests that the primary R-FGC21340 migration into Great Britain was after the R-A563 common ancestor (TMRCA in the first half of the first millennium AD), but this allows from anything from the Roman to the Norman eras. The location of the basal clades towards northern Germany suggests a post-Roman migration, with the modern English distribution covering historic Mercia and Wessex, but its bounds are likely incomplete. An Anglo-Saxon origin is posited for much of R-FGC14877. From this, we can suggest an earlier origin in the Jastorf or nearby cultures that gave rise to these Germanic tribes.

7.1.23 R-U106>Z2265>BY30097>FTT8>FGC3861>Z8053

TMRCA: Assuming 12 SNPs since R-FTT8 at a coverage of 20 Mbp, 2098 BC (95% c.i., 2475–1668 BC).

Ancient DNA:

- PCA0479, 100–300 AD, Germanic migration period, Pomerania (Poland). R-Z8053>S1855>FGC17471.
- KOS015, 650–750 AD, Merovingian Frank, Flanders (Belgium). R-Z8053>S1855>FGC17471>FGC17465>FGC17460.
- GRO012, 700–1100 AD, Frisia (Netherlands). R-Z8053>S1855(Y2404), FGC17465-.
- NTH-19, 950–1000 AD, Hungarian, Budapest. R-Z8053>S1855>FGC17471>FGC17465>FGC68720>FT153449.
- VK289, 9th Century AD, Danish Viking. R-Z8053>FGC3880.

Modern testers: 138 testers, 74 with European origins, 54 from the British Isles (34 of whom are English).

R-Z8053 contains the R-FT68373 Havilland–Verdun families, originating in Normandy, including four French and three English members, and the large historical R-FGC17467 group (40/138, 22/74, 21/54 and 17/34, the sole continental tester being Dutch).

The remainder of the haplogroup (91/138, 45/74, 30/54, 14/34) include a Frenchman, four Germans, two Danes, two Norwegians, four Swedish and three Spaniards. R-Z8053 therefore contains the only attested Scandinavian component of R-FGC3861 and, with the exception of one Dane, all fall within R-S1855.

Narrative: The basal clades of R-Z8053 include the historical Danish R-FT399057, the early medieval(?) British R-BY61970, the early medieval Spanish R-FT300525 (in Cantabria since ∼1170 AD). While Cantabria did not come under the same control of the post-Roman Germanic migrants as other parts of Iberia, it is closest to Visigothic areas, but this does not fit the distribution of R-Z8053 basal clades. The Suebi controlled areas to the west, which would more closely match the Danish and British groups, plus the overall structure of R-FGC3861 and locations of R-S1855

[p] https://www.familytreedna.com/groups/ruth/about/background

downstream. Ancient DNA is not particularly instructive during the early period, but later periods cover the west, north and east components of the Germanic expansion. A later concentration among the Germanic people appears common within R-Z8053. However, the early origins of R-Z8053 are very unclear: an origin among the Bronze Age cultures of modern Germany is suggested.

Notable sub-clades:

- R-Z8053>S1855>S1859: Contains R-FT102284, which suggests an English founder within a few centuries of 989 AD, and R-BY39524, which has a Swedish founder similiarly close to 1116 AD.
- R-Z8053>S1855>FGC17471: Contains R-FGC68720, which appears English among its three modern testers but contains ancient Hungarian DNA.
- R-Z8053>S1855>FGC17471>FGC17460: Dealt with separately below.

7.1.24 R-U106>Z2265>BY30097>FTT8>FGC3861>S1855>FGC17471>FGC17460

TMRCA: Family Tree DNA provides 198 BC (95% c.i., 638 BC – 172 AD).

Modern testers: 85 testers, 45 Europeans, 29 British.

Narrative: Exhibits a strong founder effect with rapid branching, leading to at least eight basal sub-clades. Merovingian ancient DNA from Flanders has been assigned to this group. The haplogroup's sub-clades include:

- R-BY11544, which contains the de Havilland – Verdun families of Normandy and the Battaglia family, who have been suggested to be Sicilian Vikings[25].
- R-FTC75933, which is French–Danish, containing an ancestor around 700 AD.
- R-FT115916 and R-BY122236 are (so far) uniquely German.
- R-FGC17464 (TMRCA: 108 BC – 289 AD) is older, and therefore more complex. It contains three sub-clades of its own.
  - R-BY152488 has Scots and Northern Irish testers related ∼1000 years ago.
  - R-BY115776 has Scandinavian (Norwegian, Swedish/Finnish) testers related ∼1500 years ago.
  - R-FGC17467 has mostly eastern English testers related ∼1100 years ago.

The combination of Scandinavian influences here suggests a common north Germanic theme, although it is unclear whether this applies to the common ancestor of all sub-clades.

7.1.25 R-U106>Z2265>BY30097>FTT8>FGC3861>A1243

TMRCA: Family Tree DNA provides 829 BC (95% c.i., 1395–353 BC).

Modern testers: 60 testers, 26 European testers, 20 of whom are from the British Isles. Of these, the R-BY200368 Pettit family of Suffolk comprise 33/60, 13 /26 and 11/20. The two continental Europeans in this family are French and German. There is also the R-BY45040 cluster of English families (TMRCA ∼1261 AD, 7/60, 5/26, 5/20). Together, these form R-BY45042.

The remaining R-A1243xBY45042 portion of the haplogroup comprises (20/60, 8/26, 4/20) includes two other Germans, a Pole and a Russian.

Narrative: These families do not give enough information to reliably determine an origin, but an original location somewhere in modern Germany or further east is possible.

7.1.26 R-U106 minor near-basal clades: conclusion

The distribution of modern testers, ancient DNA and projected origins for effectively all the traceable near-basal clades of R-U106 are focussed to the north-west of R-U106's posited origin of Bohemia (or nearby). Unless we have misplaced the origin of R-U106, this suggests that PNL001 represents an R-U106 individual buried as the Corded Ware Culture was still migrating westward across Europe, and that the bulk of R-U106 continued further west as it kept splitting and diversifying.

There are likely linked early paths for many of these haplogroups, which will initially have tread the same routes across Europe, but which now are hard to put together due to the passage of time. These may include haplogroups like R-FGC396 and R-S18632, which both show Bell Beaker ancient DNA in Spain. Larger haplogroups like R-FGC3861 and R-S12025 show early diversification that is difficult to piece together. R-Z18 and the now-extinct R-FGC36477 are so far the only haplogroups that show strong evidence of an early Scandinavian entry.

In summary, we can see a broad trend for early R-U106 groups to cluster in modern Germany (specifically northern Germany) and nearby. However, the detailed migration pattern for each individual sub-clade is still highly speculative and (in most cases) as much guesswork as anything else. Figure 6 shows one possible interpretation of the early R-U106 migrations, based on the above analysis.

Figure 6: A best-guess map of the migrations of R-U106 basal clades, based on their individual analysis. Dotted lines show smaller or recent migrations. This map is not expected to be entirely accurate.

Putting this in archaeological context, we can imagine that R-U106's initial expansion should follow the regions in which the Corded Ware Culture was strongest. The patterns of the subsequent haplogroups (including many of the aforemented clades during the approximate period 2600–2300 BC) will often follow the regions inhabited by the eastern Bell Beaker groups, but expansion into western Bell Beaker groups (including those in Spain, France and the UK) is likely extremely isolated and was largely unsuccessful, with these lines subsequently dying out. This leaves a nucleus of the main R-U106 population concentrated in the region of northern Germany and its surrounds.

7.2 R-U106>Z2265>BY30097>Z18

7.2.1 R-Z18 in context

TMRCA: Assuming nine SNPs since R-FTT8 at a coverage of 20 Mbp, 2284 BC (95% c.i., 2623–1897 BC).

Ancient DNA: A significant fraction of the early R-U106 burials are either directly determined to be Z18+ or are positive for other SNPs in the R-Z18 haplogroup and its sub-clades[7]. Many of these lack public information about individual calls, so it is not possible to determine whether these ancient individuals are descendants of the modern R-Z18 MRCA or not. Without this information, it is not possible to further constrain the R-Z18 TMRCA using ancient DNA.

As discussed in Section 5.5, the majority of these early burials are in Denmark, where R-Z18 sub-clades appear to trace a northern frontier to the R-U106 expansion. These pre-Migration-Age burials can be further broken down by culture (Table 4).

The concentration in Zealand is partly the result of sampling only a small number of sites. However, they establish R-Z18 in Zealand very early in the haplogroup's history. It also shows that R-Z18 was established in Sweden by the late Neolithic, and maintained a strong presence in the area through into the Iron Age.

Modern testers: 7961 testers worldwide, of which 2159 have stated European origins, of which 887 state an origin in the British Isles. The largest basal sub-clade, R-FGC79182, contains around half of this population, largely thanks to the founder effect in R-Z372 (1649 BC; 95% c.i., 2213–1160 BC).

The distribution of R-Z18 is strongly towards Nordic regions, where it is over-represented by a factor of 2.72 compared to R-U106 as a whole. Denmark is the least enhanced Scandic country (1.09), while Norway is the most (3.33). This over-representation is only clearly seen in the larger sub-clades (R-FGC79182, R-S19726 and R-CTS12023)

Table 4: Pre-Migration-Age ancient DNA from R-Z18

Sample	Date range	Culture	Region
CGG107465	2194–2026 BC	Bell Beaker	Zealand (Denmark)
CGG106705	2126–1932 BC	Nordic Bronze Age	Zealand (Denmark)
CGG106708	2125–1947 BC	Nordic Bronze Age	Zealand (Denmark)
CGG106706	∼2250–1700 BC	Nordic Bronze Age	Zealand (Denmark)
CGG105923	∼2200–1700 BC	Late Neolithic	Skåne (Sweden)
NEO752	1864–1533 BC	Nordic Bronze Age	Zealand (Denmark)
CGG106744	1730–1542 BC	Nordic Bronze Age	Langelands (Denmark)
CGG100212	1608–1430 BC	Nordic Bronze Age	Funen (Denmark)
NEO946	1322–967 BC	Nordic Bronze Age	Zealand (Denmark)
CGG100144	∼500–1 BC	pre-Roman Iron Age	Funen (Denmark)
CGG019442 (R-L257)	1–125 AD	Early Roman Iron Age	Zealand (Denmark)
CGG105930	1–150 AD	Early Roman Iron Age	Skåne (Sweden)
CGG106720	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG106722	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG106728 (R-Z372)	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG106730	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG106810	1–200 AD	Early Roman Iron Age	Jutland (Denmark)
CGG107446	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG107451	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG107494	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG107495 (R-Z372)	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG107489 (R-L257)	1–200 AD	Early Roman Iron Age	Zealand (Denmark)
CGG106489	126–227 AD	Roman Iron Age	Jutland (Denmark)

and a couple of minor sub-clades (R-BY66969, R-S7047). The remaining, smaller sub-clades tend to be more Germano–Swiss in their distribution. The occasional Meditteranean tester is also, plus a smattering from north-central Europe (especially Poland) and south-eastern Europe.

The British Isles and especially England are generally under-represented compared to R-U106 as a whole (factors 0.74 and 0.69, respectively). This is largely driven by R-FGC79182, is despite its large R-S6358 Cockburn–Dunbar cluster, which contains 274/7961, 128/2159 and 125/887 testers. The other large sub-clades, R-S19726 and R-CTS12023 show normal British Isles fractions. Among the smaller sub-clades with large enough populations, British Isles fractions are low among at least R-A6918, R-S7047 and R-BY66969 and probably others.

R-Z18 also shows a surprising concentration in Slovakia, comprising 31% (11/35) of Slovakian R-U106. Disproportionately many of these testers (9/11) are in R-Z372. A similar excess is seen in Estonia and Latvia, thanks to R-S5695.

Expansion: R-Z18 exhibits a gentle but continous acceleration of new branch formation, which traces a slow but continuous division of its sub-clades during the second millennium BC as the population grows in a stable fashion. The rate of new branch formation levels off to a stable rate of ∼9 sub-clades per century during the first millennium BC, indicating a stable population that was probably growing less rapidly. The first few centuries AD are then characterised by a rapid acceleration to ∼25 sub-clades per century, indicating another period of rapid growth, before the rate of haplogroup formation levels off again. There is no specific peak corresponding to the post-Roman Germanic migrations, as seen in many of the R-U106 basal clades.

Narrative: R-Z18 is evidently present among the most northerly outposts of the Bell Beaker Culture, in which it could originate. This feeds into its presence in the Nordic Bronze Age and modern prevalence in the Nordic countries. The lack of English testers suggests, however, that the ancestors of the Angles and Jutes nevertheless had only small R-Z18 components. Analysis of individual sub-clades is needed to form a detailed phylogeographical analysis.

7.2.2 R-Z18 minor near-basal clades

R-Z18 contains a number of sub-clades that are too small to analyse individually. These include:

- R-FT9466 (TMRCA: ∼978 BC; no testers with European origins);
- R-BY62546, an Anglo–Swiss haplogroup dating from the first millennium AD;
- R-FTA95009, an Anglo–Germano–Czech haplogroup, probably dating from Roman times;
- R-S15309, an Anglo–Germano–Belgian haplogroup, probably dating from pre-Celtic or Celtic times;
- R-S19237, a older (∼1900 BC) branch of R-Z18 whose only geographical information in its sub-clade, R-FTB22382, which comprises of two German families and one Dutch family.

At least the latter four out of these five haplogroups have locations in the southern end of both the R-Z18 and R-U106 distributions. To this group we can add several haplogroups that we can say a little more about.

R-S7047 has an expansion period beginning at R-S6133 (∼450 AD), with an Anglo–Danish–Norwegian triad of families. This timing and distribution is roughly of what we would expect from a post-Roman Germanic migration via either the Anglo–Saxon or perhaps later Viking routes. Spanish and Portuguese testers are also found via Family Finder tests, which could represent the Barbarian invasions of the Roman Empire.

R-ZP156 is the most southerly of the R-U106 haplogroups examined so far, with multiple returns from Germany, France and Switzerland, and a Migration-Age Hungarian ancient DNA result. These returns are spread well throughout its tree. Its main period of growth appears to be from about 200 AD onwards, though with enough uncertainty that this could either correspond to the haplogroup's rise being during the Roman Empire or after its fall.

R-BY66969 is a more modern haplogroup, dating from the high medieval period, and containing only Swedes and Finns. It is the most northerly of the R-U106 haplogroups examined so far.

R-A6918 contains two Scots, four Germans, three Swiss, one Pole, one Russian and one Sardinian. The Russian tester is isolated by 1500 years, but is most closely related to a German family, thus appears to represent a historical migration to Russia. The three R-ZP192 Swiss testers are related around 450 AD with no non-Swiss testers in their haplogroup, suggesting that family has been Swiss since that time. The Sardinian's R-FTD78522 has no TMRCA, but appears a Roman-era haplogroup. This leaves a spread of Germans, the Pole and the two closely related Scots. Assuming the basal R-FT115381 Scots branch is a historical immigration, this leaves a locus for this haplogroup somewhere in Germany or Poland.

This leaves the larger sub-clades, R-FGC79182, R-S19726, R-CTS12023 and R-FGC5817.

7.2.3 R-Z18>FGC5817

TMRCA: Assuming one SNP since R-Z18 at a coverage of 16 Mbp, 2166 BC (95% c.i., 2499–1788 BC).

Modern testers: 70 testers, 27 Europeans, 24 from the British Isles.

No geographical information is available until R-FGC5827 (TMRCA: 1713 BC, 95% c.i.: 2110–1269 BC, 61/70 testers). The basal clade R-FGC5827>Y292256 contains a Pole and a Portuguese. The sub-clade R-FGC5827> FGC5798>FGC5815 contains a Spaniard and the 24 British testers (it also contains a tester identifying as Israeli, but this does not look like a realistic genealogy).

Narrative: The Spaniard belongs to a deep sub-clade (R-FT55174) can could be a sporadic migration or NPE. He could also trace a deeper, untested Spanish population but, with a Spain:British bias of 12.5:1, this is less likely. Consequently we assign R-FGC5815 as possibly an entirely British sub-clade. Its TMRCA is (at 95% confidence) 79 BC – 608 AD. This covers the entire Britano–Roman period and the initial parts of the post-Roman migrations afterward, and it is suggested that the haplogroup entered Great Britain during one of these periods.

This leaves the basal sub-clades, where we have only two testers (the Pole and Portuguese) to assign an origin via. Their relationship is ancient (∼1395 BC), so it is difficult to determine a plausible origin for this group from only these data. One possibility is a Germanic group like the Suebi, but this is not contemporary, so cannot apply to their common ancestor.

7.2.4 R-Z18>CTS12023

TMRCA: Family Tree DNA provides 687 BC (95% c.i., 1178–270 BC).

Ancient DNA:

- I18184, 565–635 AD, an early Avar from Hungary.
- HAD005, 5th–6th century AD, an Angle from Cambridge. R-ZP85>FGC78525>ZP121.
- BUK005/BUK042/BUK048 ∼450–750 AD, a man of Kent. R-PH1163.
- CGG10075, medieval Danish (Randers). R-ZP85>FGC78525>ZP121.

Modern testers: 350 testers, 91 Europeans, 53 from the British Isles (of which 33 from England). These ratios are within the realms of normality for R-U106.

In north-west Europe, there are 16 testers: slightly fewer than expected (by a factor 0.74). These include two Frenchmen, ten Germans, two Swiss and two Dutch. One of the Swiss occupies the basal clade R-FT388368. There are two Poles, one Russian and one Estonian, but the haplogroup is not large enough to compare to expectations in north-eastern, eastern or south-eastern Europe, or the Mediterranean.

There is a surfeit of R-CTS12023 testers in the Nordic countries (factor 1.56). Denmark (2 testers), Sweden (8 testers) and Finland (1 tester) are close to expectations, but Norway has seven testers — around three times the expected number. This excess is confined to the R-CTS12023 basal clades, notably including R-PH1163 and R-BY73202 and a basal Swedish tester in R-S3525: the dominant sub-clade R-ZP85 does not exhibit similar Scandinavian excess.

Narrative: R-CTS12023 is relatively unique in that it covers a very wide range of northern continental Europe. Its early history is unclear, and it appears that a mixture of migrations occurred. Ancient DNA indicates that the

northern R-PH1163 sub-clade probably still extended down into the territory of the Jutes, while the presence of R-ZP85>FGC78525>ZP121 in both medieval Denmark and Angle settlements indicates a strong, long Danish presence (R-ZP85 is dealt with specifically below).

Speculatively, R-CTS12023 appears too late and too far south for the major Nordic Bronze Age groups, and too late and too far north for the major Urnfield culture. However, it is probably too early for the Jastorf culture. Some intermediate cultural package (e.g., Wessenstedt culture?) may fit better.

7.2.5 R-Z18>CTS12023>ZP85

TMRCA: Family Tree DNA provides 659 BC (95% c.i., 1167–230 BC).

Ancient DNA:

- HAD005, 5th–6th century AD, an Angle from Cambridge. R-FGC78525>ZP121.
- CGG10075, medieval Danish (Randers). R-FGC78525>ZP121.

Modern testers: 94 testers, 45 with European origins, 32 from the British Isles. The continental testers are from France (1), Germany (5), Poland (2), Denmark (1), Norway (1), Sweden (1), Finland (1) and Estonia (1). Of these testers, 74/94, 35/45 and 24/32 fall into the R-FGC78525>ZP121 sub-clade (TMRCA: ∼211 AD). The two R-ZP85xZP121 continental testers are the Dane and the Norwegian.
Conclusions: With the exception of one German tester within the recent R-FT111760 family (which we discount as a potential NPE or disputed genealogy), there are no British–continental connections after 800 AD. The British–continental connections reach a peak around 500 AD. This, and the presence of Angle ancient DNA, suggests that most of the British R-ZP85 testers are descended from the Anglo–Saxon–Jute migrations to England after the fall of Rome.

R-ZP85 remains fairly neutrally distributed around continental Europe, although 13 continental testers is not enough to determine an accurate distribution. It shows components from countries influenced by all north, east and west Germanic populations.

7.2.6 R-Z18>S19726

TMRCA: Assuming 15 SNPs since R-FTT8 at a coverage of 20 Mbp, 1863 BC (95% c.i., 2258–1425 BC).

Ancient DNA: BUK064 ∼475–750 AD, a man of Kent. R-S11601.

Modern testers: 632 testers, 174 with European origins, 100 from the British Isles, a moderately large fraction (60/100) of whom are English (cf., 51% for R-U106 overall). The majority of testers (553/632, 159/174, 93/100, 57/60) belong to the much younger R-S11601>S15815>ZP30 sub-clade, which is discussed separately.

The basal R-S19726xZP30 continental testers comprise a Frenchman, five Swedes, a Finn and a Spaniard. A tester from Flanders is also not counted in these totals. The Swedes and Finn share a common ancestor in R-BY111283 around 1100 AD. With the exception of a Scot and the tester from Flanders, the others Y-DNA results appear to be extracted from autosomal DNA tests, so are likely untyped below R-S19726.

Narrative: This therefore provides us with very little information to deduce an origin for R-S19726, as we are mostly reliant on recent haplogroups. It appears fairly typical of other R-Z18 sub-clades, with a locus near northern Germany or Denmark, but is hard to pinpoint precisely.

7.2.7 R-Z18>S19726>S11601>S15815>ZP30

TMRCA: Family Tree DNA provides 51 BC (95% c.i., 433 BC – 273 AD).

Modern testers: 553 testers, 159 Europeans, 93 of whom are British, of whom 57 (61%) are English.

The majority (47) of the continental European testers are Nordic, with an over-representation factor of 2.34 compared to R-U106 as a whole (cf., 2.72 for R-Z18 overall). There are 24 Swedes (including one basal tester), 15 Norwegians, five Danes and three Finns. Only 72 of the 159 Europeans are typed below R-ZP30 (this includes the basal Swede). The remainder are likely Family Finder or other autosomal tests.

Of the 19 non-Nordic, non-British-Isles testers, all expect one Greek are in north-western Europe, comprising ten Germans, four Dutch, one Belgian and one Frenchman. The bias-corrected locus is in the Zuider Zee in the Netherlands, but this is probably too far west for an origin due to founder effects in England.

Expansion: The haplogroup shows two main periods of expansion, between its foundation and the third century AD, and concentrated near the seventh and eighth centuries AD. There is a population contraction around the fall of Rome.

Narrative: R-ZP30 shows a relatively complex distribution for a young haplogroup. Its strong Scandinavian component persists throughout the haplogroup, but is most notable in R-ZP144>FT4479>FT4074>Y112538 (TMRCA: ∼350 AD). The medieval R-Y112538>Y95493 in particular is concentrated in Värmland (Sweden), while the late medieval R-Y112538>BY71612>BY106437 is strongly Norwegian from west of Trondheim.

Coincidentally, the R-ZP144>FT424364 haplogroup (TMRCA: ∼200 AD; 95% c.i., 171 BC – 503 AD) is entirely from the British Isles. While several of its eight sub-clades belong to medieval families, some (e.g., R-FT23542) are much earlier and still boast an entirely English cohort. The rapid branching and large number of purely British sub-clades means it is possible that R-FT424364 represents a Roman-era entry into the British Isles. Better refinement of the TMRCA through further testing is encouraged, in order to rule out either pre- or post-Roman migration, and to identify any testers that could be from outside the British Isles.

Overall, the strong Scandinavian component of R-ZP30 cannot be ignored, particularly in Norway. There is insufficient evidence to place an origin in any specific place, but an origin in Norway or Sweden might be expected.

7.2.8 R-Z18>FGC79182

TMRCA: Assuming ten SNPs since R-FTT8 at a coverage of 20 Mbp, 2225 BC (95% c.i., 2573–1830 BC).

Modern testers: 3695 modern testers, 1244 with known European origins, 494 from the British Isles. The vast majority of these belong to R-Z17, discussed separately below. The remainder belong to R-FGC72125, which comprises 43/3695, 14/1244 and 12/494 testers, the two non-British-Isles testers being French and Danish, and separate from the British testers during Roman times.

Narrative: With little data at basal levels, we cannot separate R-FGC79182 from the upstream R-Z18 and downstream R-Z17. The only additional information comes from R-FGC72125, which is old but sparsely populated, so a clear origin cannot be determined, but it does not appear Scandinavian.

7.2.9 R-Z18>FGC79182>Z17

TMRCA: Assuming 12 SNPs since R-FTT8 at a coverage of 20 Mbp, 2080 BC (95% c.i., 2442–1675 BC).

Ancient DNA: Several ancient DNA samples are specifically typed to within R-Z17>Z372, which is dealt with separately. Ancient DNA from basal R-Z17 haplogroups are:

- STR393, ∼460–530 AD, Ostrogoth?, Bavaria (Germany).
- urm160, ∼1025 AD, Swedish Viking, near Stockholm. R-S17032>BY18986>BY18987.
- GRO007, ∼985–1030 AD, Frisian, Netherlands. R-FT60052>S17721>FT111242>BY73026.
- SWG001, ∼1120–1160 AD, Jute, Schleswig-Holstein (Germany). R-FT60052>S17721.

Modern testers: 3652 modern testers, 1230 with known European origins, 482 from the British Isles. These are split into eight sub-clades, many of which are almost as old as R-Z17 itself. The vast majority (1115/1230) of testers belong to R-Z372, discussed separately below.

The remaining sub-clades show differing geographies and histories. From smallest to largest in terms of testers with known European origins:

- **R-BY40633** (13/3652, 4/1230, 0/482). TMRCA 306 AD (197 BC – 707 AD). Two Germans, two Swiss. A southern locus is inferred, but this may be historical.
- **R-BY18896** (6/3652, 5/1230, 1/482). TMRCA 1414 BC (1894 – 848 BC; basis 7 SNPs at 14.3 Mbp). German, Czech, Slovak, Danish and UK returns. A south-central locus is inferred. This southerly aspect could be ancient.
- **R-S17032** (16/3652, 9/1230, 1/482). TMRCA 1761 BC (2157 – 1305 BC; basis 7 SNPs at 14.3 Mbp). Returns from England (1), Germany (3), the Netherlands (1), Czechia (1), Norway (1), Finland (1), and Russia (1). A complex haplogroup. The sub-clade R-BY18986 (TMRCA ∼1150 BC) may be concentrated in northern and eastern Germanic regions, while the sub-clade R-S12083 (TMRCA ∼350 BC) might be concentrated in western Germanic regions.
- **R-S20045** (20/3652, 11/1230, 8/482). TMRCA 251 BC (851 BC – 234 AD). All geographical information comes from the R-S14827 sub-clade (TMRCA ∼350 AD), which has Scandinavian basal clades R-BY136156 and R-FT8378 and a lowland Scots family under R-FT4811. Probably northern Germanic in origin, with possible Viking origins for R-FT8378.
- **R-BY18864** (18/3652, 15/1230, 14/482). TMRCA 1751 BC (2147 – 1297 BC; basis 3 SNPs at 14.3 Mbp). Almost entirely British. The only continental tester is a German. Of the 14 British returns, 11 are from the R-BY18866 Dickinson family. The origin is unclear.
- **R-BY13800** (19/3652, 16/1230, 0/482). TMRCA 1241 BC (2002 – 615 BC). Almost entirely Nordic: Sweden (8), Finland (5), Norway (1). Also Germany (1) and Estonia (1). No British returns. All except the German belong to sub-clade R-BY13808 (TMRCA ∼650 AD), which appears Scandinavian in origin.

- **R-FT60052** (61/3652, 34/1230, 19/482). TMRCA 1934 BC (2282 – 1549 BC; basis 1 SNP at 16 Mbp). The non-British-Isles testers are from Germany (5), the Netherlands (4), Poland (3, closely related), Czechia (1), Denmark (1) and Sweden (1). Three of the Germans are related via R-FTA79028 (TMRCA ∼500 AD) along with an English tester. The timings of relationships between Britons and continental testers (particularly ancient DNA) suggests a Migration Age settlement in the UK, so a Danish/German/Dutch locus for this haplogroup.

Conclusions: R-Z17xZ372 still shows a wide distribution of geography, with a subset of haplogroups being strongly Scandinavian, and a subset of haplogroups showing loci around Germany. This suggests that R-Z17 was still homogeneous with the R-Z18 population, and that the population did not split until after the R-Z17 foundation.

7.2.10 R-Z18>FGC79182>Z17>Z372

TMRCA: Assuming 13 SNPs since R-FTT8 at a coverage of 20 Mbp, 1834 BC (95% c.i., 2213–1422 BC).

Ancient DNA: CGG105928, 196 BC – 218 AD, early Iron Age, Skåne (Sweden). Also others in sub-clades R-S5695 and R-Y38140.

Modern testers: 3466 modern testers, 1115 with European origins, 432 in the British Isles. These are split into the dominant R-S5695, the smaller R-Y38140, the tiny R-BY70120 (comprising two Germans related in ∼350 AD) and two basal testers (one of whom is Swiss).

R-Z372 is very strong in the Nordic countries, being over-represented by a factor of 3.50 and reaching 5.02 in Norway specifically (though it is numerically most common in Sweden). However, it is comparatively absent in Denmark (factor 0.57). This is largely down to R-S5695, but is also true to a lesser extent of R-Y38140.

R-Z372 is also strong in Slovakia, with both R-S5695 and R-Y38140 being strongly represented here.

R-Z372 is generally under-represented in the British Isles (except Scotland, thanks to the founder effect of the Cockburn–Dunbar group). It is also under-represented in north-western Europe, except Switzerland, though this is more true of R-S5695 than R-Y38140. It is also less common than other R-U106 groups in south-east Europe and the Mediterranean.

Expansion: The expansion of R-Z372 shows a relatively continuous growth, though with peaks around 900 BC and 800 AD.

Narrative: The Germano–Swiss nature of the minor basal clades indicates that R-Z372 remains within the same genetic melting pot as the overall R-Z18 population. The nature of R-Z372 is best explored through the migrations of its individual sub-clades, R-S5695 and R-Y38140.

7.2.11 R-Z18>FGC79182>Z17>Z372>Y38140

TMRCA: Assuming 14 SNPs since R-FTT8 at a coverage of 20 Mbp, 1768 BC (95% c.i., 2165–1334 BC).

Ancient DNA: R-Y38140 contains five ancient DNA results, all of which are sub-typed to R-ZP91, dealt with separately.

Modern testers: 370 modern testers, 166 with known European origins, 91 from the British Isles. Of these, most (216/370, 123/166, 55/91) belong to the immediate sub-clade R-ZP91, which represents a continuation of the initial R-Z372>Y38140 expansion.

The remaining R-Y38140xZ372 testers are mostly British (154/370, 43/166, 36/91). There are also four Swedes and three Frenchmen. One basal R-Y38140>BY41647 tester is Swedish, the others are R-Y38140>CTS5860. One of the Swedes is typed to the medieval R-CTS5860>S4037>BY1285. The others remain untyped.

The British R-Y38140xZ372 testers largely fall into three medieval families: the Scottish R-BY1285 Nesbitt family (Ayrshire), the south-west English R-S3315, and the Scots–Irish R-BY20396 Young family. There is also the smaller, younger R-FT236738 group, who lack a European origin.

7.2.12 R-Z18>FGC79182>Z17>Z372>Y38140>ZP91

TMRCA: Assuming 15 SNPs since R-FTT8 at a coverage of 20 Mbp, 1712 BC (95% c.i., 2117 – 1270 BC).

Ancient DNA: SZ 4, ∼550–570 AD, Langobard, Hungary. R-Y98441>FT423338. Also four more under R-BY41788.

Modern testers: 216 modern testers, 123 with known European origins, 55 from the British Isles. Overall, R-ZP91 appears strong in the Nordic countries (over-representation factor 1.61) and north-central Europe (2.80), particularly the Czech Republic (4.85) and Austria (6.80). Its bias-corrected median position is near the tripoint between the Czech Republic, Germany and Poland.

Narrative: The haplogroup consists of nine sub-clades, six of which subsequently split during the middle Bronze Age. Two of the remainder are medieval English families, while the final haplogroup (R-FTA95555) is Classical-Age Anglo–Dutch.

Of the older small haplogroups:

- R-FT20270 is Germano–Finnish;

- R-BY120333 has returns from Norway, the Netherlands and a medieval family from the Russian state of Karachay–Cherkessia in the Caucasus;
- R-BY211482 is Brittano–Swedish; and
- R-Y98441 is Germano–Brittano–Swedish, and also contains the Langobard SZ 4.

These groups appear to be a mix of Germanic groups, and while R-ZP91 does not have the significantly southern component of upstream haplogroups, they are stronger in north-central Europe. The family from Karachay–Cherkessia is particularly intriguing, but its interpretation is unclear. This leaves the larger R-S5970 and R-BY41788, which are dealt with separately, below. These haplogroups continue the mix of Alpine and Scandinavian components, with indications of a migration both north and south in the final millennium BC. Note that the Lombards are mentioned in the discussion on R-BY41788, as well as being represented here in ancient DNA.

7.2.13 R-Z18>FGC79182>Z17>Z372>Y38140>ZP91>S5970

TMRCA: Family Tree DNA provides 1169 BC (95% c.i., 1828–618 BC).

Modern testers: 83 modern testers, 44 with European origins, 20 from the British Isles. The 24 European testers comprise men from Germany (6), the Netherlands (3), Austria (2), Denmark (1), Norway (7), Sweden (3), Bulgaria (1) and Portugual (1).

Narrative: This is a difficult haplogroup to assess, due to the mix of countries and migrations that have occurred over the last ∼3000 years. There are minor components like the high medieval Anglo–Swedish R-FT37207, which we can naïvely treat as Viking in origin, and the Germano–Dutch R-FT407000 and R-BY99947, which could be up to 2000 years old. The only sub-clade with a strikingly clear origin is the Roman-era / early medieval R-FT76491, which is clearly Norwegian. This haplogroup therefore has a complex migration history to both the north and south that requires more testers to uncover.

7.2.14 R-Z18>FGC79182>Z17>Z372>Y38140>ZP91>BY41788

TMRCA: Assuming 15 SNPs since R-FTT8 at a coverage of 20 Mbp, 1664 BC (95% c.i., 2085–1203 BC).

Ancient DNA:

- STR316. ∼480–510 AD. Ostrogoth(?) from Bavaria. R-ZP136.
- KOS032. ∼650–750 AD. Merovingian from Flanders. R-S7015>BY19948>BY71305.
- VDP-A7. ∼850–1050 AD. Early Icelandic. R-S7015>BY19948>BY71305>FT209682.
- ATP_PSN_496. ∼1300–1400 AD. Medieval Cambridge. Pre-R-S7015>BY172778>FTB18868.

Modern testers: 100 modern testers, 54 with European origins, 21 from the British Isles. The European testers comprise men from France (1), Germany (10), Switzerland (2), Poland (1), Czechia (6), Slovakia (1), Austria (2), Sweden (8), Russia (1) and Italy (1). The testers divide slightly unequally into R-S7015 and R-ZP136. Both have very strong returns in the area of the Czech Republic and surrounds; both contain German and Swiss testers.

A notable shift occurs in R-ZP136>BY84754 (25/100, 16/54, 0/21; TMRCA: 382 BC [964 BC – 94 AD]), which concentrates in north-western (6/16) and north-central (8/16) Europe. Its sub-clade R-BY121244 contains almost all of the north-central European testers, while its other sub-clades contain the north-western European testers. The other R-ZP136 clade, R-FT259169, is strongly Swedish (with one English tester).

R-S7015 provides fewer clues. The Swiss–Swedish haplogroup R-BY80372 possibly dates to the Migration Age, but the dates are unclear.

Conclusion: The disparity between R-ZP136>BY84754 and R-ZP136>FT259169 indicates a migration occurring some time (perhaps shortly) after the R-ZP136 foundation (TMRCA: 821 BC, 95% c.i.: 1463–290 BC). This involved both migrations north to (or perhaps within) Sweden and south to both the western Germanic regions of the Franks and the eastern Germanic regions of the Goths. This likely establishes R-ZP136 and, by extension, the wider R-BY41788 as a pre-Germanic haplogroup, as its components appear to have been involved in the early peopling of the Germanic world during the final centuries BC. A migration from the Elbe river to Bohemia and Italy would be consistent with the history of the Lombards, but there is insufficient evidence to attach any particular Germanic tribe to this group.

7.2.15 R-Z18>FGC79182>Z17>Z372>S5695

TMRCA: Assuming 14 SNPs since R-FTT8 at a coverage of 20 Mbp, 1788 BC (95% c.i., 2172–1372 BC).

Ancient DNA: 11 ancient DNA results, split between R-L257 (5) and R-S4031 (6). The earliest of these are from Iron Age Denmark.

Modern testers: 2759 modern testers, 882 with European origins, 326 from the British Isles. Of these, a significant fraction belong to the R-L257≫S6358 Cockburn–Dunbar cluster, which contains 274/2759, 128/882 and 125/326 testers. R-S5695 represents 35% of R-Z18 testers overall.

Geographically, the haplogroup shows a highly significant over-representation in the Nordic countries (by a factor of 3.94). Topologically, it splits into the slightly larger R-L257 and slightly smaller R-S4031, with a third basal clade, R-FTB9467, comprising of only two historically related Czech testers.
Narrative: R-S5695 formed as part of the continuing R-Z372 expansion. Since little evidence can be obtained from R-FTB9467, it is best understood as the amalgam of its two major sub-clades.

7.2.16 R-Z18>FGC79182>Z17>Z372>S5695>L257

TMRCA: Assuming 15 SNPs since R-FTT8 at a coverage of 20 Mbp, 1667 BC (95% c.i., 2060–1246 BC).

Ancient DNA: Five ancient DNA samples. Four belong to the sub-clade R-Z8185>Z15. The final sample is KOS006 (650–750 AD), a Merovingian Frank from Flanders, typed as R-FT417873>FTF25483.

Modern testers: 1438 modern testers, 419 with known European origins, 279 from the British Isles. The vast majority belong to R-Z8185, and the vast majority of those belong to R-Z15. A significant portion of the R-L257 testers are assigned based on autosomal tests, and no assignment is made below R-L257.

Overall, the haplogroup shows a strong concentration in north-central Europe, particularly Slovakia, where it is over-represented by a factor of ∼9. This is based on only eight testers, but is nevertheless a significant result. It has a much more normal presence in the Nordic countries (factor 0.83) than much of R-Z18. However, like other R-Z18 haplogroups, there are fewer returns than average from north-western Europe (0.65) and eastern Europe (0.76), south-eastern Europe (0.48) and the Meditteranean (0.27). Its presence in the British Isles is above average, thanks to the R-Z15>S6358 Cockburn–Dunbar cluster: removing this cluster gives a fairly average fraction of testers from the British Isles.

Considering the basal clades of R-L257 individually:

- The small R-FT417873 contains individuals from Germany and Northern Ireland, plus KOS006. We can clearly see Germanic influence here, but there is insufficient information to determine a distribution.
- R-FGC69963 is modern Germano–Swiss, and presumably follows the same distribution as many of the other Swiss-strong haplogroups under R-Z18.
- R-FT75149 is an older and diverse branch of R-L257, containing a basal Italian tester, a Viking Age Finnish family, and the medieval R-ZP198 Dunn cluster.
- R-Z8185>FT58372 contains little information. Aside from a basal Polish tester, there are two medieval English and one Irish families who share a common ancestor in R-FT58177 (∼300 AD).
- R-Z8185>B314 is more interesting. It splits into two major sub-clades. R-BY172021 contains a basal Polish tester and a high medieval Slovakian family with representatives in the Ukraine and Poland. R-BY155840 contains a basal English tester and R-BY66127, a wholly Swiss haplogroup dating to around 700 AD.

Conclusions: There is little information in the small basal clades to distinguish R-L257 as a whole from R-Z8185 or even R-Z15. It is difficult to determine whether the southerly focus of R-L257 as a whole is due to its founder having migrated south or the preferential survival of its southern migrant lines. The presence of clear Nordic groups within R-Z8185>Z15 suggests the latter.

We can therefore interpret R-FGC69963 and particularly R-B314 as migrations to southern Germanic countries. The timing of this migration is expected to be between the oldest likely age for R-B314 (1444 BC at 84% confidence) and the youngest likely age for R-BY155840 and R-BY172021 (217 BC and 33 BC, respectively).

7.2.17 R-Z18>FGC79182>Z17>Z372>S5695>L257>Z8185>Z15

TMRCA: Family Tree DNA provides 1385 BC (95% c.i., 1942–905 BC).

Ancient DNA: Four samples:

- DUN011, 672–773 AD, Frisian from Lower Saxony, R-S23346>S11880.
- SH-175, 950–1000 AD, Hungarian conqueror, R-ZP141>FT96427.
- VK204, 900–1000 AD, Viking in Orkney, R-ZP141>BY93324>BY115469.
- VK308, 900-1150 AD, Swedish Viking, Västra Göthaland, R-Z378>BY33037.

Modern testers: 801 modern testers, 274 with known European origins, 209 from the British Isles. The Cockburn–Dunbar group comprises 274/801, 128/274 and 125/209 of these. Considering the remainder, these are distributed around Europe in a very typical way for a R-U106 haplogroup. There is possibly a slight concentration towards north-central Europe. However, these numbers may be affected by autosomal testers whose tests are only as detailed as R-Z15.

The four basal clades of R-Z15 are all considerably younger than R-Z15 itself, suggesting a rekindling of a dying haplogroup. They are:

- R-FT21888 contains a Polish tester and the British Elder family, related around 2000 years ago.
- R-ZP141 is diverse. Its smaller clades contains a Swede, a Lithuanian and the Hungarian SH-175, suggesting an eastern Germanic component. R-BY93324 contains a German and a Dutch tester, related in early medieval times, but also contains the strongly Norwegian R-BY115469, which VK204 indicates has Viking origins.
- R-S23346 contains Swedish testers in both of its sub-clades, however it is also found in the Netherlands (cf., DUN011), suggesting a complex migration after its formation.
- R-Z378 splits into the much larger R-Z375 and the smaller R-BY33037. The latter contains at least some Viking ancestry (VK308) but also has a Dutch haplogroup, R-FTD19873, which is over 1000 years old.

Narrative: R-Z15 has a complex migration pattern that is not easily understood via its living testers or ancient DNA. R-Z15 shows the same mix of Nordic and southern Germanic groups as the much of the rest of R-Z18, suggesting it remained part of the same, fairly homogeneous population between the R-Z18 founder and the R-Z15 founder, despite the ∼1000 year difference between them.

7.2.18 R-Z18>FGC79182>Z17>Z372>S5695>L257>Z8185>Z15>Z378>Z375

TMRCA: Family Tree DNA provides 195 BC (95% c.i., 573 BC – 129 AD).

Modern testers: 687 modern testers, 241 with known European origins, 201 from the British Isles. The Cockburn–Dunbar group comprises 274/687, 128/241 and 125/201 of these. Considering the remainder, these are still distributed around Europe in a very typical way for a R-U106 haplogroup, with a possible slight concentration towards north-central Europe. However, these numbers may be affected by autosomal testers whose tests are only as detailed as R-Z375.

R-Z375 presents a founder effect, with three main sub-clades:

- R-A15148 contains a basal Estonian tester. A German tester is indicated, but is very closely related to an English tester and has English Y-STR matches, suggesting possible misattributed ancestry.
- R-Y27977 exhibits a rich sub-structure, showing continual growth over the past 2000 years and very varied geography. Its basal testers are generally distributed around Denmark, Norway and Sweden. The sub-clade R-FTB42644 appears uniquely British, but it is not large enough to confirm it has a British-dominated distribution. The sub-clade R-ZP204 (circa 300–900 AD) shows a mostly continental distribution, covering Germanic countries, and ranging from the Netherlands to Poland while including both Norway and the Orkney islands. Much of this migration appears historical, complicating efforts to assign an origin.
- R-ZP8 is dealt with separately below.

Narrative: The complex distribution of R-Y27977 and consequent reliance on the analysis of R-ZP8 prevents a clear picture of the movement of the R-Z375. However, the general alignment with the Germanic countries and the dynamic nature of the haplogroup corresponds well with the major expansions of the Germanic peoples.

7.2.19 R-Z18>FGC79182>Z17>Z372>S5695>L257>Z8185>Z15>Z378>Z375>ZP8

TMRCA: Family Tree DNA provides 112 BC (95% c.i., 484 BC – 206 AD).

Modern testers: 395 modern testers, 173 with known European origins, 161 from the British Isles. The Cockburn–Dunbar group comprises 274/395, 128/173 and 125/161 of these, but they are not the only historical family to be over-represented. The 16th/17th Century English R-FTA79136 family comprises 18/395, 15/173 and 15/161 testers, while the R-S4052 Allen family comprise 11/395, 4/173 and 4/161. This leaves only 92/395, 26/173 and 17/161 testers outwith these families.

Excepting the Cockburn–Dunbar group, the majority of the British families in this group are English. The basal clades of R-ZP8 contain testers from the Ukraine, the Netherlands, Germany (R-FTC74990) and France (R-A15898).

The dominant R-FGC45254 sub-clade represents a continuation of the initial expansion. Its basal clade R-A11482 contains another Frenchman, plus the aforementioned English R-FTA79136 and Cockburn–Dunbar precursor group, R-ZP2.

Conclusions: R-ZP8 appears much more of a western Germanic group than the haplogroups above it. The Frenchmen near the root of the haplogroup are relatively unusual, and probably indicate a relatively strong French component, hidden by the lack of testing in France. However, the presence of the Ukranian basal tester suggests that the cut from the other Germanic groups wasn't complete by this point, and that migration within the Germanic world still occurred at significant enough levels to detect.

The Dunbar family (and the Cockburns by association) are supposed to be descended from Crinan of Dunkeld (975—1045). They are not the only family to make this claim, but have the best pedigree to prove it, descending from his grandson Gospatric, Earl of Northumbria (d. 1073). The date of the Cockburn–Dunbar split is unknown, but probably dates from the 1100s, and is commensurate with R-S5750 (TMRCA: 1069 AD; 858–1243 AD). The R-S6358

haplogroup likely predates this relationship by about 200 years, and probably represents this haplogroup's entry into the British Isles.

The R-S6358xS5750 families include both Scottish and English families, though few Irish. They include the R-A20777 Clan Rutherford, who originate in Roxburghshire in the 1100s. The majority of the other families are distributed up the east coasts of England and Scotland, from Yorkshire to Aberdeenshire. The close proximity of the Dunbar and Rutherford lands suggests an origin for R-S6358 in the north of medieval Northumbria. Its arrival here could have been either during Anglo–Saxon or Viking times. The west Germanic leanings of R-ZP8 overall suggests the latter is more likely, though a Danish Viking origin is also easily possible.

7.2.20 R-Z18>FGC79182>Z17>Z372>S5695>S4031

TMRCA: Family Tree DNA provides 1269 BC (95% c.i., 1832–794 BC).

Ancient DNA: Six samples within R-S3207>CTS5533.

Modern testers: 1319 modern testers, 461 with European origins, 47 from the British Isles, 396 from the Nordic countries. Vastly dominated by the younger sub-clade R-S3207.

Narrative: R-S4031xS3207 only contains useful geographical information for one modern tester from Scotland. R-S4031 must be understood solely in terms of R-S3207.

7.2.21 R-Z18>FGC79182>Z17>Z372>S5695>S4031>S3207

TMRCA: Family Tree DNA provides 974 BC (95% c.i., 1481–539 BC).

Ancient DNA: Six samples within R-CTS5533.

Modern testers: 1291 modern testers, 457 with European origins, 46 from the British Isles, 395 from the Nordic countries (121 from Norway, 244 from Sweden, 23 from Finland).

Conclusions: R-S3207 is a very strongly Scandinavian haplogroup, with the large majority of its testers being Norwegian or Swedish. This extends to all three sub-clades: R-CTS5533, R-S5673 and even the tiny R-BY63337. Of these, only R-CTS5533 forms part of the initial R-S3207 expansion. However, we can interpret these results as a migration to the Scandinavian peninsula (probably Sweden) around the TMRCA of R-S3207.

Within R-S3207, the median location of the Norwegian and Swedish results lie between Oslo and Stockholm, with R-CTS5533 being slightly further east than R-S5673. Sub-clades of both R-CTS5533 and R-S5673 generally concentrate in the north of Götaland and south of Svealand, especially around the Swedish lakes Vänern and Vättern where the homeland of the Geats (Goths) is situated. Only R-CTS5533 has a substantial Finnish population.

7.2.22 R-Z18>FGC79182>Z17>Z372>S5695>S4031>S3207>S5673

TMRCA: Family Tree DNA provides 281 BC (95% c.i., 720 BC – 90 AD).

Modern testers: 662 modern testers, 219 with European origins, 14 from the British Isles, 200 from the Nordic countries (68 from Norway, 128 from Sweden). All except two testers without geographic origins belong to R-S5684 (TMRCA ∼100 BC).

Conclusions: R-S5684 is very strongly Swedish, with the exception of a few Norwegian groups.

The sub-clade R-S5686 provides an interesting case where five Scottish testers are included in a Norwegian group. The Scottish testers are from Sutherland and Caithness, and are presumably descended from the Viking-era settlement of these areas (∼900–1098 AD). The sub-clade R-S5686>BY106715 becomes Norwegian again later on, indicating back migration to Norway.

R-BY873 represents the larger part of R-S5684, and breaks down into the strongly Swedish R-FT80702 and mixed Swedish–Norwegian R-ZP108. The basal clades of R-ZP108 are mostly Norwegian, while the Viking Age sub-clade R-FGC36338 is Swedish, while the slightly earlier R-BY12550 is Norwegian. These likely trace a series of minor migrations around the two countries, but there is insufficient geographic data shared to ascertain an exact route.

7.2.23 R-Z18>FGC79182>Z17>Z372>S5695>S4031>S3207>CTS5533

TMRCA: Family Tree DNA provides 861 BC (95% c.i., 1378 – 422 BC).

Ancient DNA:

- VK418, 300–400 AD, Iron Age northern Norway, Y20021+? (pre-R-CTS2158>S6989).
- VK170, ∼950 AD, Isle of Man, R-CTS2158>S6989>S3201.
- VK449, 980–1009 AD, Danelaw Dorset (England), R-FT20255>FT22694.
- VK259, 980–1009 AD, Danelaw Dorset (England), R-FT20255>FT22694.
- gam872, 950–1100 AD, Viking from Uppland (Sweden), R-BY19581>FT10809>BY19580>Y42202.

- kro001, 1616–1676 AD, *Kronan* (Swedish warship), R-FT20255>FT22694.

Modern testers: 572 modern testers, 229 with European origins, 30 from the British Isles, 189 from the Nordic countries (53 from Norway, 111 from Sweden). Five Germans, one Czech, a Hungarian, two Latvians, one Italian.

R-CTS5533 contains five sub-clades:

- One basal tester of unknown origin.
- R-FTB47774, containing a single medieval Scottish family.
- R-FT25359, shows a clear Viking presence through ancient DNA and modern Swedish/Norwegian testers, but also Czech and Hungarian individuals suggesting a southern Germanic component too.
- R-BY19581: the basal clades begin with a significant Swedish component, though also contain Scottish, German and Finnish testers. The early medieval sub-clade R-FT10809>BY19580 splits into the R-BY62045 McLeod family of Caithness and the high medieval Finnish R-Y42202>Y43130, both of which appear to be Viking in origin (cf., gam872).
- R-CTS2158 forms the dominant part of R-CTS5533. Apart from a basal Finnish tester, all results are within R-S6989.

Conclusions: Relatively little migration out of the Scandinavian peninsula is seen in R-CTS5533, though it is apparent that some back migration to continental Europe and expansion of R-CTS5533 among the Germanic groups occurred. However, there are too few returns to identify which particular route or group was involved. A clear Viking origin for individuals in the British Isles is seen.

7.2.24 R-Z18>FGC79182>Z17>Z372>S5695>S4031>S3207>CTS5533>S6989

TMRCA: Family Tree DNA provides 388 BC (95% c.i., 851 BC – 2 AD).

Ancient DNA: VK170, ∼950 AD, Isle of Man, R-S3201.

Modern testers: 346 modern testers, 165 Europeans: 17 British/Irish, 142 Nordic (39 Norwegian, 89 Swedish), 3 Germans, 2 Latvians and an Italian.

Narrative: R-S6989 has six sub-clades, all of which date from a few centuries after its foundation (nominally 100 BC – 150 AD). The Latvian and a Northern Irish tester are basal, the Italian–Danish sub-clade R-Z19523 is the most southerly of the sub-clades. R-S3201 and R-A14188 are largely Swedish, while R-BY27836, R-BY16535 and R-Y30157 are more of a Norwegian–Swedish mix. Patterns do exist in the data that can separate the migrations of individual families within R-S6989, but this is not possible with the available public data.

7.2.25 R-Z18 conclusions

The putative migrations discussed above are encoded in Figure 7. While the exact locations and migration directions of each clade are not expected to be accurate in most cases, the map gives a sense of the complexity involved, and some of the generalised migration directions that can be anticipated.

The key feature of R-Z18's geographical distribution is its northerly aspect. It has the most northerly midpoint of all the major R-U106 haplogroups and the strongest Scandinavian presence, both in ancient and modern DNA. Simultaneously, R-Z18 is over-represented in many countries along the southern and eastern borders of R-U106's distribution. Examination of individual haplogroups within R-Z18 do not show uniquely Scandinavian or uniquely southern/eastern haplogroups until relatively close to the present, indicating that both the Scandinavian and Alpine distributions of R-Z18 are the result of a relatively late migration. Consequently, in placing the R-Z18 origin, it was decided to adopt a middle ground, with the origin of R-Z18 somewhere in the vicinity of the mouth of the Elbe and the Danish marches.

It is worth remembering that Figure 7, as well as likely hosting many inaccuracies, only shows the backbone of ancestry that leads to the major groups of modern testers. Thus, even if the MRCA of R-Z18 was born in this region, it does not preclude extinct R-Z18 lines living nearby, such as the early individuals whose ancient DNA has been found in northern Jutland. An origin in the south of Jutland is used instead as we do not see much travel to and from Norway and Sweden in the early stages of R-Z18's growth. However, the placement of haplogroups on the map should not be treated as correct to that level of precision.

Archeologically, this location is close to the southern boundary of the Nordic Bronze Age. To account for CGG107465, we have to have some method of getting R-Z18 into the Bell Beaker culture, without having it take part in the main R-P312 thrust along the Atlantic coasts. For the map in Figure 7, a route is taken from the Corded Ware Culture of Bohemia along the Elbe. This sticks in the Corded Ware Culture main territories, but settles in the Mittelelbe–Saale component of the Bell Beaker culture, which would allow an easy move from the mouth of the Elbe up into the Jutland Bell Beaker component. However, this is only one of many options.

From this origin, we can then trace the various migrations to the north and south. Dating these migrations accurately has proven to be difficult. It is likely that there was more than one north–south migration involved. However,

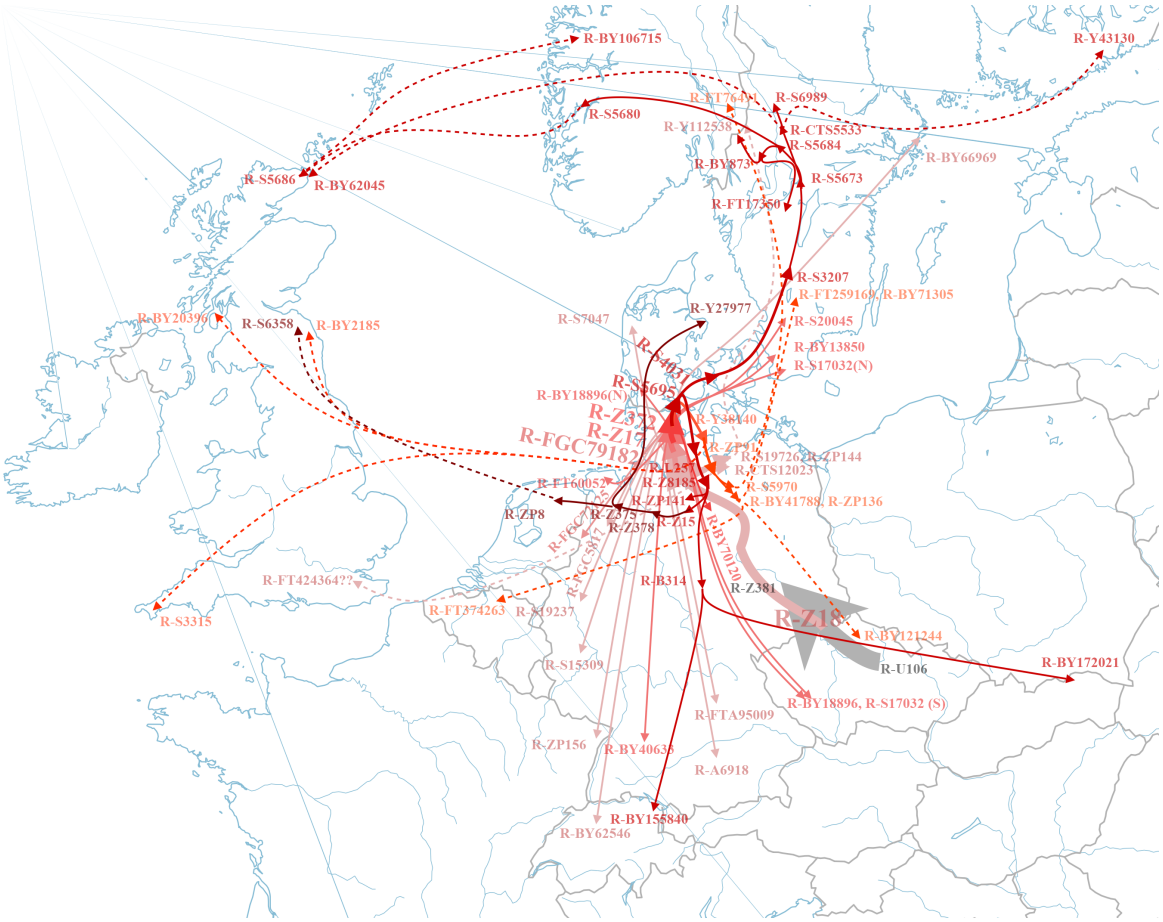

Figure 7: A best-guess map of the migrations of R-Z18 basal clades, based on their individual analysis. Dotted lines show smaller or recent migrations. This map is not expected to be entirely accurate.

we would normally expect one migration to dominate numerically among these, and the TMRCAs of upstream and downstream haplogroups can be used to bracket the date of this dominant migration.

Within R-Z372, the best timing for the dominant migration probably comes from the haplogroups R-ZP136>BY84754, which is fully Alpine, and R-ZP136>FT259169>ZP155, which is fully Swedish. The TMRCA for R-ZP136 provides an older limit to the migration of 1141 BC (at 84% confidence), while the TMRCA for R-BY84754 and R-ZP155 provides a younger limit of 134 BC. Similarly, R-B314 and its subclades suggest a date between 1444 BC and 217 BC. This period ties in with the expansion of the early Germanic peoples, which would provide a natural route to take a migration south from Jutland and the Danish marches down into Alpine Europe.

Similarly, we can date the dominant migration north to Scandinavia using R-S5695 and R-S3207, whose TMRCAs give a probable range (at 84% confidence) of 1902–752 BC. This migration could therefore also be associated with the rise of the Germanic peoples, but could also easily be before that, during the earlier phases of the Nordic Bronze Age.

A scattering of eastern Europeans exist in many R-Z18 haplogroups. None of these eastern haplogroups are of sufficient size to say much about their origins, history and migrations. However, migrations into the Wielbark Culture that led to the Gothic peoples have potential to explain some of these isolated cases. Alternatively, these could be later Viking off-shoots, or even the result of traders in the Hanseatic League.

Overall, it appears likely that most R-Z18 families passed through the Germanic sphere of influence at some point, and R-Z18 was likely part of the Germanic peoples from their very beginning. While R-Z18 is too old to be given the title of a Germanic haplogroup, R-Z18 testers should consider a Germanic history as very likely for their individual families.

7.3 R-U106>Z2265>BY30097>Z381 minor near-basal clades

This section excludes R-Z156, R-L48 and R-S1688, which are discussed separately. This section is based on the haplotree as of 2025 February 17.

7.3.1 R-Z381 in context

TMRCA: Assuming one SNP since R-FTT8 at a coverage of 20 Mbp, 2752 BC (95% c.i., 3046–2398 BC).

Jinonice I7196 (Section 5.5.5) can be used to constrain this age very slightly. An age of 2075 ± 125 BC (at 68% confidence) is adopted. It is not possible to assign a precise percentage possibility that the S1911+ and S1894+ reads

of I7196 are correct. We assume a 50% probability in both cases, and assume that the three reads for Z304+ provide a 100% probability. This means that we attribute a 100% chance of I7196 being Z304+, a 75% chance of S1911+ and a 50% chance of S1894+. A coverage of 20 Mbp between R-FTT8 and R-S1894 is also assumed: in the case of lower coverage, I7196 limits the age of haplogroups between R-Z381 and R-S1894 further.

Z304 lies nine SNPs below R-Z381. However, S1911 and S1894 (respectively) have two and three equivalents in R-S1911 and R-S1894. This means that S1911 could lie 12, 13 or 14 SNPs below R-Z381 and S1894 could lie 15, 16, 17 or 18 SNPs below R-Z381. This gives a 25% probability of constraint nine SNP below R-Z381, an 8.3% probability of constraint 12, 13 and 14 SNPs below R-Z381, and a 12.5% probability of constraint 15, 16, 17 and 18 SNPs below R-Z381. This provides substantial constraint in the TMRCAs of sub-clades down the lineage of I7196, but for R-Z381 only restricts the TMRCA to 2791 BC (95% c.i., 3063–2536 BC).

Ancient DNA: Few early geographical constraints on R-Z381 exist in ancient DNA. The earliest sample is I7196 (Jinonice, Prague; Únetiče culture; R-Z381>Z156≫Z304?S1911?S1894). Other early samples are found in the Elp culture, from its earliest phases (I4070; north Holland; 1880–1657 BC) to its height (I11972; north Holland; 1501–1310 BC). Later results are found in the Hilversum, Urnfield and Celtic cultures (see R-Z156).

CGG106838 (2281–2048 BC, Zealand, Denmark; Bell Beaker) has been reported to be R-Z381>Z301>FGC13959>S9891, but this age is inconsistent with the TMRCA of this haplogroup (900 BC). The individual calls from this sample have not yet been released, so it is possible that CGG106838 is R-FGC13959 and S9891, but not R-S9891. In this case, CGG106838 could be additional minor constraint on the age of R-Z381. However, if S9891 is simply a bad call, we do not know what the next upstream SNP is, so we do not include it in the analysis here.

Modern testers: R-Z381 men make up 84% of R-U106 testers, with R-Z18 making up the majority of the remainder. It is therefore not easily possible to differentiate R-Z381 from the remaining bulk of R-U106. The haplogroup appears to represent approximately 12% of the British Isles, 14% of north-western Europe, 5% of north-central Europe, 5% of the Nordic countries, 2% of eastern Europe, 2% of south-eastern Europe, 2% of the Meditteranean countries, 9% of European populations overall and 8.9% of the testers at Family Tree DNA, specifically.

There are two major haplogroups, R-Z301 and R-Z156. R-Z156 has its own section, R-Z301 is discussed next. There are also two minor haplogroups: R-FT40367 and R-M323. Little can be said about these minor haplogroups because they contain only British testers. Their arrival into the British Isles is unclear, but could be consistently explained with a series of early medieval migrations (i.e., "Anglo-Saxons" to Normans). R-M323>BY20775 in particular is strongly Welsh and probably pre-dates the Norman conquest. A basal Italian tester also exists.

Expansion: R-Z381 shows a relatively rapid diversification in its R-Z156 and R-Z301 sub-clades and, with five immediate sub-clades itself, can still be considered as being part of the initial major R-U106 population expansion.

Narrative: From the TMRCA, it can realistically be expected that R-Z381 was still part of the initial Corded Ware Culture expansion. Here, R-Z381's origin is (mostly arbitrarily) placed in the middle German part of the culture's field of influence, on the Elbe river for the sole reason that the Elbe gives the best access from Bohemia to the regions where R-Z381 is common.

The clear presence of R-Z381 in the Únetiče culture of Bohemia a few centuries after its foundation could mean it arose there and only migrated out of Bohemia later. However, the much stronger presence of all major R-Z381 sub-clades in western Europe indicates a bulk R-Z381 migration west of Bohemia rather than any other direction. The middle Elbe is also close to the western bounds of the Únetiče culture, so not inconsistent with the adoption of R-Z381 into Únetiče practices.

Given its size at the time, R-Z381 is suprisingly absent from the Bell Beaker groups (with the probable exception of CGG106838 in Denmark). This, and the lack of participation in the R-P312 migrations (as traced by a lack of geographical overlap with R-P312) is another factor in keeping R-Z381's origin further east.

7.3.2 R-Z381>Z301

TMRCA: Assuming two SNPs since R-FTT8 at a coverage of 20 Mbp, 2693 BC (95% c.i., 3003–2311 BC).

Ancient DNA: Aside from CGG106838 (see R-Z381, above), there are no other Bronze Age ancient DNA results for R-Z301. If confirmed R-Z301, this shows that R-Z301 entered the Danish Bell Beaker group with R-Z18. However, R-Z301 does not show the same Scandinavian-dominated distribution today as R-Z18 does, likely indicating the bulk of the haplogroup remained further south.

Modern testers: R-Z301 comprises 77% of the European R-Z381 testers and 64% of the European R-U106 testers at Family Tree DNA. This makes it difficult to separate the distribution of R-Z301 from R-U106 as a whole. R-Z301 shows a lower Scandinavian fraction than R-U106 as a whole, but consistent with the removal of R-Z18 from the comparison pool. Of the remaining sub-clades, R-S10807 has only one English tester, so we cannot say anything further about it. R-FGC20667, R-FGC13959 and R-FGC8512 have their own sections, below.

Expansion: R-Z301 has six sub-clades. The two largest, R-L48 and R-S1688, have their own sections. Apart from R-S1688, the sub-clades are separated from R-Z301 by only 1–2 SNPs, indicating a continued relatively rapid expansion, although perhaps not quite at the scale of the early R-U106 through R-FTT8 expansion.

Narrative: R-Z301 is indistinguishable in distribution from R-Z381 as a whole, and shares much of the same distribution as R-Z156 and R-Z18: with the exception of the strong Scandinavian of R-Z18, the distributions of these three major

haplogroups appear very similar, so it's likely that R-Z301 and R-Z381 formed in the same population. We have therefore indicated a very similar location for the origin of R-Z301 and R-Z381.

7.3.3 R-Z381>Z301>FGC13959

TMRCA: Assuming two SNPs since R-Z301 at a coverage of 16 Mbp, and allowing constraint from CGG106838, 2500 BC (95% c.i., 2853–2189 BC).

Ancient DNA: CGG106838 (2281-2048 BC; Zealand, Denmark; Bell Beaker) represents the only prehistoric ancient DNA in R-FGC13959. See R-Z301 and R-Z381 for further detail, including its unclear placement in the haplotree.

Historical-era ancient DNA is found in

- early medieval Hungary (Karos 3-13; 895–950 AD; R-BY41605 and Árnád 55; c. 700–900 AD; R-BY41605>BY13391>FT333 and
- late medieval Germany (Petersberg 820; 1020–1116 AD; R-BY41605>A7222).

Modern testers: 342 testers, 67 Europeans, 15 from the British Isles. The British Isles fraction is much less than typical for R-U106 (40% of normal).

Continental testers are strongly found in north-central Europe, particularly in the Czech Republic, and in the Mediterranean, especially Portugal.

The haplogroup splits into two sub-clades, R-S9891 and R-BY11543.

Expansion: R-FGC13959 shows very little expansion until about 1000 BC, when a slow but constant expansion began on both sub-clades.

Conclusions: Germany, the Czech Republic and a few British Isles results are common to both sub-clades. The Portuguese and Spanish results are only found in R-BY11543≫BY41605>BY133391, and R-BY41605 also contains an Estonian tester. R-S9891 contains Norwegian and Swedish testers.

The Czech testers in R-S9891≫FTC16026 are related during the first millenium AD, setting the latest timescale at which their mutation could have happened. This could easily be coincident with the Magyar migrations, given the ancient DNA results (though these do not appear to be from eastern Europe or the Urals where the Huns are supposed to originate).

The group's only Scandinavian testers, R-S9891≫BY68937 are related between the fourth and 14th Centuries AD, with the highest probability around the Viking Age.

R-BY41605 arose during the time of the rise of both Germanic and Celtic groups. There is a strong German component in this group, but it also contains the Estonian tester and others. One tester with a German country designation is actually from Silesia in Poland. It's not clear to which ethnic group R-BY41605 belongs, but an origin in the Hallstadt-era Celts could be considered.

Particularly, the R-BY41605≫FT333761 group contains R-BY41605's Spanish and Czech testers. Given the other Magyar-era Hungarian burial in this group, there seems to be a central European – Iberian link in this particular sub-clade, which dates to between ∼1000 BC and ∼500 AD. The Spanish connection could be relatively recent, medieval or classical era.

Meanwhile, the R-BY41605≫FT367218 Portuguese connection is Roman-era or medieval, as the two Portuguese testers share a common ancestor in this timescale. This could represent migration within the Roman empire, or a post-Roman Germanic migration. If post-Roman, then the Suebi migrations might best fit, with an origin in Germany, and a diaspora in Portugal, the Czech Republic and Hungary. The Vandals offer another option to get families in Spain. However, these are only two of many possible interpretations.

7.3.4 R-Z381>Z301>FGC20667

TMRCA: Assuming two SNPs since R-Z301 at a coverage of 16 Mbp, 2473 BC (95% c.i., 2845–1982 BC).

Ancient DNA: Two modern-era ancient DNA results: Hofstaðir 128 (940∼1070; Iceland; R-FGC23826>Y96503) and Vor Frue Kirkegård 445 (1500s; Aalborg, Denmark; R-FGC23826>FTE63927).

Modern testers: 104 testers, 61 Europeans, 31 British Isles. These are unequally distributed between two much younger haplogroups, R-FGC20676 and R-FGC23826.

Expansion: The expansion of this haplogroup is very slow, but contains a rapid peak in the Viking Age.

Conclusions: R-FGC20676 dates to the middle to late Bronze Age. It contains one basal clade (R-FGC20669) containing an English and an unknown tester. The remainder belongs to the Scottish R-A14208 family which dates from circa 800–1250 AD. This comes from the Scots (rather than Gaelic) half of Scotland, south of the Highland Line, and predominantly from the east coast. Their origins before Scotland are unclear.

R-FGC23826 dates to the Iron Age. It contains five immediate sub-clades. R-FTB39753 is a colonial era Smith family. R-FT362693 is a medieval family of possible British origin. The remaining three (R-Y96503, R-FTE63927 and R-Y19258) appear to be Iron Age Scandinavian, setting a likely origin for the entire haplogroup. A mix of Danish and Swedish ancestry means it is unclear which of these countries is the origin. R-Y19258>Y18877>Y18884 in particular appears to be Viking Age Norwegian.

7.3.5 R-Z381>Z301>FGC8512

TMRCA: Assuming one SNP since R-Z301 at a coverage of 16 Mbp, 2547 BC (95% c.i., 2898–2089 BC).

Ancient DNA: R-FGC8512 contains two early medieval ancient DNA samples, both in R-BY3251>Z155.

Modern testers: 577 testers, 167 with known European origins, 99 from the British Isles. Of the continental Europeans, there is a relatively strong Meditteranean component (5/68) spread over multiple countries and clades. A medieval Estonian family occupy R-Z155≫BY41246 (7/68). There is a relatively small Scandinavian component (10/68), mostly Swedes (5) and Danes (4). Finally, there is a fairly typical concentration in north-western Europe (41/68), including Germans (25), Dutch (6) and a sizeable population of French (9).

Expansion: R-FGC8512 represents the end of the initial period of expansion for this line. Its two descendant sub-clades start to expand again around 1250 BC and continue a slow expansion over the subsequent millennia.

Narrative: Without early branching, it is hard to determine the origin of R-FGC8512. Of its sub-clades, the older R-BY3251 is too diverse and too sparsely populated to have a clear point of origin or migration pattern. It has two testers from Italy or Malta. This may not be a statistically significant contribution, and it is not clear when this migration took place or from where. Otherwise, it contains a fairly normal mix for R-U106. The larger R-Z155 is dealt with separately, below.

7.3.6 R-Z381>Z301>FGC8512>Z155

TMRCA: Family Tree DNA provides a TMRCA of 1230 BC (95% c.i., 1790–750 BC).

Ancient DNA: Two ancient burials: Dunum 6 (600∼1000 AD) is a Frisian from Lower Saxony and R-Z153. Sedgeford 3 (667–773 AD) is an Angle-era burial from Norfolk and R-Z153>Z363>Z154.

Modern testers: 538 testers, 155 with European origins, 94 from the British Isles.

Many of these testers belong to a number of heavily tested historical families. Within R-BY19015, there is the Mumma family (R-FGC8507), a branch of the Momma family of Aachen, another of whom moved to Stockholm and was enobled as Reenstierna. The wider R-BY19016 family appear to be of German (or Dutch–German) descent, dating back to between the late Bronze and Roman ages.

Basal testers of the other basal sub-clade, R-Z153, include a Dutch tester and the aforementioned Frisian ancient DNA sample. There is a potential Iberian (Canary Isles, Brazil) conncetion in R-Z153>FGC49658 that dates back to the late Bronze or Iron Ages, but this is unclear.

The sole basal tester of R-Z153>Z363 is Swedish. R-Z363 also contains the circa Norman-era, mostly Scottish R-BY41247, and R-Z154, which is discussed separately below.

Narrative: The origins and spread of R-Z155 are difficult to tease apart. It is clearly very strong in north-western Europe, particularly around the North Sea coasts. This could point it an origin and homeland, but this can't be confidently stated at present.

7.3.7 R-Z381>Z301>FGC8512>Z155>Z363>S3503>Z154

TMRCA: Family Tree DNA provides a TMRCA of 206 AD (95% c.i., 148 BC – 504 AD).

Ancient DNA: Sedgeford 3 (667–773 AD) is an Angle-era burial from Norfolk.

Modern testers: 228 modern testers, 67 Europeans, 47 from the British Isles. The British Isles testers are mostly English (35/47). The continental testers are French (3), German (3), various Scandinavian (4), Greek (1), Russian (2) and Estonian (7).

There are several historical families among these. These include:

- R-Y31452>Y31448 Sweetster of Herefordshire (14/228, 13/67, 13/47, 13/35);
- R-Y31452>BY64941 Walker of Lancashire (7/228, 6/67, 6/47, 6/35);
- R-Z356>BY65758 Warner (18/228, 7/67, 5/47, 3/35, 1 German, 1 French);
- R-Z356>BY19022>BY41246 of Estonia (4/228, 4/67, presuambly plus three Family Finder testers); and
- R-Z356>BY19022>BY60904 Gann (12/228, 1/267, 1/47, 1/35).

Expansion: R-Z154 experienced an initial period of rapid expansion, probably during Roman times, before a hiatus and a later expansion during early to high medieval times.

Narrative: Despite being a relatively populous haplogroup, the concentration of individuals in well-tested families means that there are relatively few independent continental testers to use to assign origins.

At face value, the family could be Angle in origin, which would explain the presence of the ancient DNA result in Norfolk, the nearby ancient Sweetster family in Herefordshire, and many of the other English testers. This is less likely to be the case if the haplogroup is on the older end of the TMRCA.

The Estonian R-BY41246 family is harder to diagnose. It predates the Hanseattic league but post-dates the split of the Germanic peoples. One possibility is that it is the result of the Northern Crusades.

Figure 8: A best-guess map of the migrations of R-Z381 basal clades, based on their individual analysis. Dotted lines show smaller or recent migrations. This map is not expected to be entirely accurate.

7.3.8 R-Z381 basal clades: conclusions

Figure 8 shows a cartographic representation of the possible migrations discussed above. In many ways, these minor clades, centuries removed from the Corded Ware Culture foundation, are some of the hardest haplogroups to define phylogenetic origins for.

That said, these haplogroups have (on the whole) relatively little to distinguish them from the R-U106 bulk until we examine smaller sub-clades closer to the present. This suggests are relative homogeneity among the R-U106 populations during these early phases. Consequently, while we have placed an origin for R-U106 in Bohemia, the majority of these minor clades are shown on Figure 8 as following the same well-paved route as much of the rest of R-U106, north-west from Bohemia. The necessity of this is dictated by the presence of R-U106 in both ancient and modern DNA to the north-west, and by the lack of R-U106 in other directions.

7.4 R-U106>Z2265>BY30097>Z156 minor near-basal clades

This section is based on the haplotree as of 2025 February 17.

7.4.1 R-Z156 in context

TMRCA: Assuming two SNPs (including Z8160) since R-FTT8 at a coverage of 20 Mbp, 2634 BC (95% c.i., 2959–2226 BC). Constraint from Jinonice I7196 (Section 5.5.5) can be used to constrain this age slightly using the same principles as for R-Z381. Similarly, we imply a 25% probability of constraint seven SNPs below R-Z156, an 8.3% probability of constraint 10, 11 and 12 SNPs below R-Z156, and a 12.5% probability of constraint 13, 14, 15 and 16 SNPs below R-Z156. This provides moderate constraint in the TMRCAs of sub-clades down the lineage of I7196 and, for R-Z156 specifically, it restricts the TMRCA to 2702 BC (95% c.i., 2987–2461 BC).

Ancient DNA: A total of 33 ancient DNA results are found in R-Z156. Aside from I7196 (Jinonice, Prague; Únetiče culture; R-Z381>Z156≫Z304?S1911?S1894), the other significant ancient DNA result is I11149, significant because of its age and location (Teversham, Cambridgeshire; Iron Age Britain; 733–397 BC). Family Tree DNA assigns I11149 only as R-Z156, but re-analysis with a T2T reference sequence[q] records Z5889+ (7 reads), with mixed or no reads at other SNPs and a single possible read for FT186928 (the burial is inconsistent with the TMRCA of R-FT186928). While

[q] `https://groups.io/g/R1b-U106/message/6769`

the call at Z5889 appears reasonably secure, the lack of intervening positive SNPs means we retain the designation R-Z156?Z5889.

Modern testers: R-Z156 men make up 19% of R-U106 testers. They therefore make up a sizeable fraction of R-Z156, but one small enough that we can compare R-Z156 against R-U106xZ156 as a whole.

R-Z156 is distributed fairly similarly to other R-U106 groups, with some exceptions. It shows a normal fraction of British testers (1701/2950 = 58%, cf., 56% for R-U106 as a whole). However, compared to other R-U106 groups, these are more common in Ireland (particularly Northern Ireland) and Scotland, and less common in England.

In north-western Europe, it shows a much higher fraction in France (by a factor 1.88) than the rest of R-U106. This occurs across R-U106 sub-clades. It also shows a slightly higher than average concentration in Belgium, Germany and Switzerland, and a slightly lower than average concentration in the Netherlands. This pattern persists across all R-U106 sub-clades, except perhaps for R-S5520, which is perhaps less common than usual in France.

R-Z156 is only sporadically present over eastern Europe, with the notable exception of R-DF98, which is twice as common as the rest of R-U106 in the Czech Republic, and R-S5520, which has a normal frequency over the region.

R-Z156 is comparatively absent in the Nordic Countries, with only 56% of the usual R-U106 frequency (78% of the usual R-Z381 frequency). This fraction becomes lower once the dominant R-Z304 sub-clade is ignored, reaching only 31–46% of the usual R-U106 frequencies for R-BY20378, R-S3311, R-FGC39801 and R-S5520.

The relative absence continues in eastern Europe (44% of R-U106 frequency), but R-Z156 shows typical frequencies in south-eastern Europe and (again apart from R-S5520) slightly higher than average frequencies (factor 1.45) the Meditteranean.

The distribution of R-Z156 reaches a peak in Belgium (4.3% of the testing population) and the Netherlands (3.5%), declining to 2.8% in Germany and France, 2.3% in the British Isles and 2.1% in Denmark. In the south, it maintains 0.6% across Portugal, Spain and Italy. In the Scandinavian peninsula, it makes up 0.9% of the testing population, declining to 0.4% in Finland. In the east, it makes up 1.3% of the Czech Republic, 1.1% of Austria and Hungary, 0.7% in Slovakia, 0.5% in Poland, declining to 0.1–0.2% in countries in the south-east and east of Europe.

Eight sub-clades are known, including one basal Scottish tester.

- R-A9590 contains two testers, Portuguese and English, related around 2150 BC.
- R-Y30585 contains ten testers related around 2200 BC. These are split into two sub-clades: R-Y29891 contains a historically related Finn and Estonian; R-BY61869 has a common ancestor around 1600 BC, but six of its eight testers belong to R-FT4917, related around 950 BC, including a Slovakian, a Scot and an Irishman.
- R-BY20378, R-S3311, R-FGC39801, R-S5520 and R-Z306 represent its other sub-clades, in ascending order of size, and have their own sections below.

Expansion: The expansion of R-Z156 continues the rapid expansion of the haplogroups above it, so can be thought of as part of the initial expansion of R-U106. This expansion dies down over the first few centuries, until it picks up again around 1700 BC and continues unabated until around 700 BC. Following 700 BC, the rate of haplogroup formation drops significantly, picking up again in the last few centuries BC. (Note that these dates assume a TMRCA for R-Z156 that is approximately 100 years later than the estimate above.)

R-Z306 is by far the dominant sub-clade today. However, in the initial few centuries of growth, R-S3311, R-S5520 and R-Z306 all expanded at similar rates.

Narrative: With some forgivable, the sub-clades of R-Z156 offer a remarkable geographical coherency, with their mean latitude further south and perhaps west than the rest of R-U106. The ancient DNA of R-Z156 is also found further south and often further west (Section 5.5.9), with R-Z156 being the first R-U106 haplogroup found in the British Isles (Section 5.5.2). R-Z156 is therefore fundamentally different from the other haplogroups, either by virtue of staying where it was or migrating to somewhere different. The homogeneity of its sub-clades shows that the migration was at the R-Z156 level, not below. We should therefore seek to place the origin of R-Z156 further south and perhaps west of the R-U106xZ156 haplogroups.

Given PNL001 and I11149, it is tempting leave the origin of all haplogroups between R-U106 and (at least) R-Z304 in Bohemia, and it would certainly appear reasonable to do so. However, frequentist arguments make this less likely, since R-Z156 is almost universally found west of Bohemia. While the overall distribution of R-U106 north-west of Bohemia can be explained by its foundation during a north-westerly migration, if R-U106 through R-Z304 remain in Bohemia, the same argument cannot be used for R-Z156.

Conversely, R-Z156 cannot have migrated too far south or west, otherwise it would have been taken up in the R-P312 migrations, especially the R-U152 migrations of the upper Rhine and Danube, thus be found more commonly in France and Italy. (R-U152 being a rough contemporary of R-Z156). The origin of R-Z156 must also be close enough that it can quickly be brought back into the Únětice culture to create I11149. Very speculatively, this leaves us with a swathe of central to southern Germany in which to place the migration of R-Z156.

7.4.2 R-Z156>BY20378

TMRCA: Family Tree DNA provides 1569 BC (95% c.i., 2277–975 BC).

Modern testers: 66 testers, 24 with European origins, five from the British Isles.

A notably small British component, the haplogroup is most common in Germany (8 testers), but also found in France, Belgium, the Netherlands, Austria, Finland and Portugal.

The haplogroup splits into R-BY30254 and R-FT41641. Both of these are middle Bronze Age groups, but show very different distributions.

Expansion: The haplogroup expands generally continuously between its foundation and modern times. A slight increase in haplogroup formation around 500 AD is possible.

Narrative: R-BY30254 shows a group that is predominantly German, with the odd Dutchman, Beglian and Austrian thrown in. A general absence of genealogical information among these Germans makes it difficult to isolate any particular part of Germany as its locus.

R-FT41641 is an altogether different beast. It has an early spilt into two sub-clades of its own. R-BY49885 contains English, German, French and Finnish families without connections between them since the Bronze Age. R-BY33328, and particularly its late-Bronze-Age / Iron-Age sub-clade R-BY33333 are different.

R-BY33333 splits into two haplogroups, R-FT413423 and R-BY33335. Both are likely Roman or early medieval in age. R-FT413423 is Austro-German, while R-BY33335 has two basal testers from Spanish colonies and a slightly younger group of Portuguese. Most of the Portuguese come from the Azores, but one family is from northern Portugal. This could represent migration, presumably from modern Austria/Germany to Iberia, during the Celtic, Roman or post-Roman eras.

7.4.3 R-Z156>S3311

TMRCA: Based on 16 Mbp coverage and one, five, six and ten SNPs below R-Z156:

- R-S3311: 2560 BC (95% c.i., 2884–2223 BC)
- R-FT8323: 2269 BC (95% c.i., 2671–1774 BC)
- R-Y3965: 2196 BC (95% c.i., 2617–1667 BC)
- R-S3995: 1899 BC (95% c.i., 2396–1256 BC)

Ancient DNA: Colmar 239 (740–390 BC; middle-aged La Tène B Celt; Aude, France; R-FT8323>Y3965>S3995>BY20561>A106

Modern testers: 156 modern testers, 34 with known European origins, 26 from the British Isles. The British testers include 11 English and eight Welsh. This large Welsh fraction is split among at least two clades from opposite sides of R-S3311: the 1000-year-old R-S3997 and the much younger R-BY114023. R-S3997 contains roughly half the haplogroup's testers (70/156, 10/34, 9/26) so represents a significant founder effect. R-BY114023 contains a tester claiming to be descended from the Jones of Monmouthshire (founder: Caradog ap Rhiwallon, b.∼1000), but another family in R-L96 makes the same claim.

R-S3311 is strong particularly in France, with four testers plus the ancient DNA test. The geographical locus of R-S3311 lies in France. Its remaining continental testers are found in Germany, Denmark, Sweden and Italy (Sicily).

Expansion: No clear expansion pattern is seen, given the small size of the haplogroup.

Conclusions: R-S3311 consists largely of a main line (R-S3311>FT8323>Y3965>S3995) with small basal branches at each level. At some point between R-Z156 and R-Y3965, its geography appears to have shifted significantly, possibly indicating a migration to France (though that distinction is based on a single Breton R-Y3965 tester). Better geographical information from French and German testers would assist a more precise analysis.

In any case, R-S3311 must have been established in France by the La Tène B period, and R-BY20126 provides a more definite French component that probably predates the La Tène Celts.

R-S3997 has clearly been in Wales for around 1000 years, however the TMRCA is not sufficiently accurate to date it as pre-Norman. It may have arrived anywhere from pre-Celtic times (1500 BC) to after the Norman invasion.

7.4.4 R-Z156>FGC39801

TMRCA: Based on two SNPs below R-Z156 and 16 Mbp coverage, 2488 BC (95% c.i., 2831–2107 BC).

Ancient DNA: Two Roman-era individuals, both detailed under R-FGC39800.

Modern testers: 146 modern testers, 77 with European origins, 39 from the British Isles. It splits into two sub-clades, R-FGC39800 and R-A9555.

R-FGC39801 is very strong in north-western Europe (30 testers; over-representation factor 1.6). This is mostly due to a much higher factor in the sub-clade R-A9555. Representatives are also found in Poland, Denmark, Sweden, Russia, Greece and Spain. In the UK it is strongest in Ireland, due to representation in R-FGC39800.

Expansion: R-FGC39801 shows an initial period of growth that peaks during the middle Bronze Age. There is a haitus in branch formation centred in the few centuries around 500 BC, which likely indicates a population contraction. The haplogroup then enters a period of slow and sporadic growth, with a peak around 500 AD. This pattern is seen in both sub-clades.

Narrative: The geographical differences between R-FGC39800 and R-A9555 mean they need to be understood separately before the ancestry of R-FGC39801 can be put together.

7.4.5 R-Z156>FGC39801>FGC39800

TMRCA: Based on seven SNPs below R-Z156 and 16 Mbp coverage, 2122 BC (95% c.i., 2562–1563 BC).

Ancient DNA:

- R10659; newborn; 26–126 AD; Klosterneuburg, near Vienna; Roman-era Pannonia; R-FGC39815>BY126375.
- R11121; late Roman Empire (1–400 AD); Isola Sacra, near Rome; R-BY125277.

Modern testers: 68 modern testers, 28 with European origins, 16 from the British Isles. The 12 non-British-Isles testers are from Germany (6), France, Belgium, Denmark, Sweden and Greece.

R-FGC39800 splits geographically into three parts.

1. R-BY125277 contains the Roman R11121 and the historical R-BY103288, including the American R-FTC17522 Weaver family and the likely English R-BY104277.
2. R-FGC39815>FGC39806>A7172 is a middle Bronze Age haplogroup that contains most of the haplogroup's British testers and few others.
3. R-FGC39815xFGC39806 is largely found in the Germanic countries, especially Germany itself, and tends to avoid the British Isles.

Expansion: See R-FGC39801.

Conclusions: There is relatively little evidence of R-U106 in the Roman Empire. That we see two R-FGC39800 individuals is remarkable, and speaks of substantial Romanisation of this haplogroup. It requires that a significant fraction of R-FGC39800 existed south of the Rhine–Danube limes during the Roman conquest.

This makes it possible that R-BY103288, sub-clades of R-A7172 and particularly R-BY126375>BY55726 are Roman-era arrivals to the UK. However, the haplogroups are insufficiently large to be clear on this point and they could also easily be later arrivals (e.g., post-Roman Germanic or Norman).

The widespread nature of R-FGC39815 identifies a well-spread haplogroup, but one apparently remaining largely within the Germanic sphere of influence. Hence, this is the assignment we provide to R-FGC39800 overall: while its existence in the Bronze and Iron Ages is unclear, by the Roman era, this group appears Germanised, with some families entering into the Roman sphere of influence and migrating onward from there to Italy, Greece and beyond.

7.4.6 R-Z156>FGC39801>A9555

TMRCA: Based on 13 SNPs below R-Z156 at 16 Mbp coverage and applying the Family Tree DNA TMRCA of 1934 BC (95% c.i., 2672–1310 BC) as a constraint, 1676 BC (2227–957 BC).

Modern testers: 78 modern testers, 49 with European origins, 23 from the British Isles. Isolated European testers exist in Spain, Russia, Sweden and Poland. However, the bulk (22) of the European testers come from north-western Europe (over-representation factor 1.89). Most (14) are German, but the strongest over-representation is in France (3.3), Belgium (2.7) and the Netherlands (2.8). This is therefore a relatively localised haplogroup for its age.

Expansion: See R-FGC39801. R-A9555 in particular shows rapid initial branching.

Conclusions: Given the rapid branching of R-A9555, it makes sense to look down these branches in turn and look at the basal branches to find migratory patterns. The basal R-BY17936 clade is strongly French, and only about 300 years younger than R-A9555 itself. This probably indicates that this clade migrated to France during this intervening 300 years.

R-FGC66753>BY36061 appears Germano-British, while R-FGC66753>FGC66737 has a more cosmopolitan distribution, retaining German influence throughout. The strong Irish presence among the British Isles testers is largely down to R-BY36031 group, which is likely a 1000-year-old Irish family. This age could indicate a Norman origin, though earlier migrations (e.g., early Christian migrations) cannot be ruled out.

7.4.7 R-Z156>S5520

TMRCA: Based on two SNPs (S5520, FGC11701) below R-Z156 and 16 Mbp coverage, 2488 BC (95% c.i., 2831–2107 BC).

Ancient DNA: I23978; infant, 742–400 BC, Hallstadt C or D; Zagorje ob Savi, Slovenia; R-FT221936.

Modern testers: 721 testers, 266 with European origins, 191 from the British Isles. Its subclades, in order of increasing size, are R-FTG7579, R-FT117057, R-FT221936, R-FGC48296 and R-FGC11662. Of these, only R-FGC48296 and R-FGC11662 are of reasonable size.

The haplogroup is strongly biased by the R-FGC11674 MacMillan/McMullen group. This 1000-year-old family, of Scottish origin, comprises a substantial fraction of R-S5520 (173/721, 112/266, 111/191) and the majority of R-FGC11662. While R-S5520 as a whole is dramatically biased to Scotland and Ireland, removing this group means R-FGC11662 and R-FGC48296 share similar distributions.

The MacMillan family are not the only sizeable British-Isles family that could skew the analysis. The Irish R-FGC11662>BY16554≫BY35111 family (25/721, 12/266, 12/191) is of similar age. The Scottish R-FGC48296≫BY99021 family (9/721, 7/266, 7/191) may be slightly younger and concentrates roughly in Perthshire. The Irish R-A9845 family is also around 1000 years old and is brother to the Welsh R-FT222212 family (together in R-Y19785, 12/721, 7/266, 7/191).

Even accounting for these Scottish, Irish and Welsh groups, R-S5520 is starkly absent from England, with a representation only 52% of a typical R-U106 haplogroup (28% before these groups are removed).

While R-S5520 has some Nordic representatives, they are few in number (68% of normal, mostly R-FT117057 and R-FGC48296). There are no known eastern Europeans. Instead, R-S5520 appears common in the centre of the continent, being common in Switzerland, Belgium and Germany (NW Europe: 1.6 times normal) and north-central Europe (2.0 times normal). Sporadic results in south-eastern Europe, including I23978, indicate that R-FGC11662 could also be about three times R-U106's normal frequency there too.

Considering the basal clades:

- R-FTG7579 contains one German.
- R-FT117057 contains English, Swedish, Danish and German testers.
- R-FT221936 contains the Slovenian ancient DNA result and the historic British R-FT222217 family.

Expansion: A strong start, then reasonably continuous throughout history. A rise around 400 or 500 AD is seen, possibly related to the post-Roman Germanic migrations. A possible minimum is seen around 100 BC, which could be due to the demise of the Celtic groups at the hands of either the Germans in the north or the Romans in the south.

Narrative: The similarity of R-FGC11662xFGC11674 and R-FGC48296 suggests we can treat R-S5520 as a single, homogeneous unit that later migrates to different places. This makes it one of the best tracers of the overall R-Z156 origin.

The age of the Scots, Welsh and Irish groups is consistently around 1000 years, inviting the possibility that these are Norman settlers. However, the lack of English testers suggests that R-S5520 did not participate strongly in the Anglo-Saxon, Danelaw or Norman invasions that predominantly affected England. The lack of Scandinavian groups means we can probably rule out Norse Vikings as an origin as well. So it is conceivable that these Scots/Welsh/Irish families were already in the British Isles during Roman times, and their absence in England and consequent marginalisation to the archepeligo's extremities comes as a result of these later migrations into England. There is very little information constrain their earliest arrival into the British Isles, the earliest reasonable date being the middle Bronze Age (the age of R-FGC11662).

A small component (R-FT117057, or maybe just its sub-clade R-FGC9217) does appear to have migrated to Scandinavia at some point in the first few centuries of R-S5520's growth. This could be related to the Bell Beaker migration in that direction, thus tied to R-Z18's fate. Other Scandinavian testers are consistent with small-scale migration within the Germanic world.

Many of the near-basal testers (R-FTG7579, R-FGC11662*, etc.) appear German or from neighbouring countries (France, Poland) but avoiding the Low Countries and Denmark, suggests an overall origin of R-S5520 in southern Germany, consistent with R-Z156 overall. A large number of unlisted origins in some parts of R-S5520 make it difficult to work out the migration patterns of intermediate (middle/late Bronze Age) haplogroups. However, there are a few more recent examples which appear more homogeneous in their country of origin:.

- R-FGC11662≫BY16559 is 2500 years old and strongly German with some Swiss and Polish.
- R-FTA25007 is 3000–2500 years old and strongly German with some Irish. This may extend as far as R-FGC48296>FTA17325, which is 4000 years old and has a basal Swiss tester.
- R-FGC48296>S5556>BY33288 is strongly German, though its sub-clades include a 1000-year-old southern Italian group R-FTA22046, and the 1000–1500 year-old Czech-strong group R-BY33291.

The Germano–Irish connection in particular suggests a Celtic connection was important to R-S5520 in the last millennium BC.

7.4.8 R-Z156>Z306

TMRCA: Assuming six SNPs since R-FTT8 at 20 Mbp coverage, 2397 BC (95% c.i., 2781–1898 BC). Adding constraints from I7196, this becomes 2544 BC (95% c.i., 2842–2351 BC).

Ancient DNA: There are 28 ancient DNA results under R-Z306, of which I7196 is the oldest. Family Tree DNA places three of them at the R-Z306 level:

- 3DRIF-16, which can be typed to R-Z307>Z3044≫DF96≫L1;
- AED 106, a Migration-Age Bavarian; and
- I11574, a Wessex-era burial in Worth Matravers, Dorset.

The latter two ancient individuals are likely R-Z304 too.

Modern testers: R-Z306 contains the majority (75%) of R-Z156 European testers and a sizeable fraction (15%) of R-U106 overall. This is largely thanks to its sub-clade R-Z307>Z304, which makes up the vast majority of its testers, detailed analaysis of which is left to that haplogroup.

R-Z156xZ307 is comprised entirely of R-BY41601, a middle Bronze Age haplogroup comprising 16 members of whom at least eight are from the British Isles, with no other European members. The Iron Age sub-clade R-BY41905 splits them into the probably post-Norman Anglo-Welsh R-BY67531 and the probably pre-Norman Welsh R-BY50734.

Narrative: Little can be made of the early days of R-Z306, as all information either comes from R-Z307 or a very late British branch. The arrival of R-BY41601 in the British Isles is debatable, and could be any time from the Bronze Age to (possibly) the Norman conquest.

7.4.9 R-Z156>Z306>Z307

TMRCA: Assuming eight SNPs since R-FTT8 at 20 Mbp coverage, 2278 BC (95% c.i., 2691–1739 BC). Adding constraints from I7196, this becomes 2472 BC (95% c.i., 2772–2301 BC).

Ancient DNA: All 26 ancient DNA results under R-Z307 are in R-Z304.

Modern testers: R-Z307 represents the vast majority of R-Z306 and is comprised almost exclusively of R-Z304. R-Z307xZ304 is comprised entirely of R-FGC44894, an early Bronze Age haplogroup comprising the historical-era French R-FTE36660 family and the 1000-year-old probably Scottish R-BY3237 family.

Narrative: Again, little can be made of the early history of R-Z307, with information either coming from R-Z304 or the tiny R-FGC44894. The French component is interesting, as it suggests there could be a substantial untested French branch within the basal clades of R-Z307 and maybe R-Z306.

7.4.10 R-Z156>Z306>Z307>Z304

TMRCA: Assuming nine SNPs since R-FTT8 at 20 Mbp coverage, 2218 BC (95% c.i., 2646–1660 BC). Adding constraints from I7196, this becomes 2437 BC (95% c.i., 2737–2275 BC).

Ancient DNA: There are 26 ancient DNA results under R-Z304. Of these, five have only been sequenced to the R-Z304 level:

- I17019; 1421–1216 BC; Hilversum culture of south Holland.
- I13788; 1300–800 BC; Urnfield culture of north-west Bohemia.
- I12907; 356–57 BC; Iron Age in north Holland.
- BUK027; 475∼750 AD; Jute in post-Roman Kent.
- DUN010; 600∼1000 AD; Frisian in Lower Saxony.

Along with I7196, I17019 and I13788 represent the first three Z304+ burials.

Modern testers: 9148 modern testers, 2225 with European origins, 1295 from the British Isles. R-Z304 represents 14% of R-U106 testers in Europe. It contains two significant sub-clades: R-BY12480>FGC8365>DF96 (1101/2225) and R-FGC29253>DF98 (704/2225), which make up the majority of its testers.

Minor basal clades are:

- R-BY60581, a Celtic- or Roman-age haplogroup containing five individuals, including two Irish testers and an Austrian;
- R-FT19354, a Roman- or post-Roman-age haplogroup containing a German and a Czech; and
- R-BY167599, a late Bronze Age or early Iron Age haplogroup containing a basal French tester and an Iron-Age pair from Switzerland and Germany.

R-FGC29253xDF98 comprises R-Y71329. This is a Bronze Age haplogroup containing a basal Belgian and a Roman/post-Roman German family.

R-BY12480xFGC8365 contains R-BY12482 while R-FGC8365xDF96 contains R-A10971. These are significant enough to be given their own section, below.

The frequency of R-Z304 reaches a maximum in Belgium (3.3%) and the Netherlands (2.9%), decreasing towards Germany (2.2%) and France (2.1%). In the British Isles, it is most common in England (2.4%) but remains common in Scotland (1.4%) and Wales (1.2%). In the island of Ireland, the frequency drops to 1.2%, but is about a factor of three more common in Northern Ireland than the Republic.

In the north, R-Z304 remains common in Denmark (1.6%) but drops in the Scandinavian peninsula to 0.7%. In the south, it remains common among Alpine countries: Switzerland (1.5%) and Czechia (1.1%) especially.

Compared to other R-U106 haplogroups, R-Z304 has a typical fraction in the British Isles, in which (apart from Northern Ireland) it is also distributed like most other R-U106 haplogroups. It is slightly (factor 1.15) more common in north-west Europe, especially in Belgium (factor 1.34, thanks to R-DF98) and France (factor 1.82). R-DF98 is also common in the Meditteranean, though R-DF96 is not. R-Z304 has a normal fraction in south-east Europe.

R-Z304 is comparatively absent from other parts of Europe. This includes north-central Europe (68% as common as other R-U106 groups), with the exception of R-DF98 in the Czech Republic. R-Z304 is also rare in Scandinavia (63% of normal) with the exception of Finland (127% of normal). R-Z304 is very rare in eastern Europe (48% of normal), especially R-DF98.

The bias-corrected geographical locus of R-Z304 lies in the German Palatinate, near the French border. This is slightly north-east of the overall R-Z156 locus. The R-DF98 locus is in the Moselle valley in Lorraine, while the R-DF96 locus is east of Frankfurt.

Expansion: Ignoring R-DF98 and R-DF96, the near-basal clades of R-Z304 show a relatively uniform expansion across pre-Classical and Classical history. A possible rise occurs around 500 AD.

Narrative: It is again tempting to place the origin of R-Z304 in Bohemia and let the R-U106 (PNL001) to R-Z304 or later (I7196) thread of ancient DNA carry on in Bohemia unabated. This is certainly possible, but only if all the post-Únětice migration out of Bohemia was towards the west, as this is where the majority of the R-Z304 population is today. Furthermore, it is really only the R-DF98 component of R-Z304 that has a meaningful presence in the Czech Republic. Herein, the origin of R-Z304 is therefore placed in southern Germany, where most of the other R-Z156 haplogroups have been placed, although its exact origin is uncertain.

While the expansion from the Corded Ware culture to R-Z304 was relatively continuous (only a small break between R-Z156 and R-Z306 shows), R-Z304 nevertheless represents a turning point where the population expanded substantially. In the chronology discussed here, this would best correspond to the rise of the Únětice culture.

Whether R-Z304 lay at the heart of the Únětice culture or perhaps more towards (or beyond) its western boundaries can be debated. On the one hand, we have I7196 firmly embedded in the early phases of the culture, so we know that at least some of the haplogroup was involved in it. On the other hand, the modern distribution of testers is heavily skewed towards regions further west, even after testing biases have been corrected for. When the Únětice culture disbanded around 1700 BC, several migrations formed an exodus. There are three circumstances we can consider that might still place R-Z304 in the centre of the Únětice culture in Bohemia:

1. Either migrations out of the Únětice culture were lop-sided, with most people going west, or migrations west were ultimately more successful and those in the east died/daughtered out.

2. R-Z304 happened to be predominantly in the westward migrations (so can't have been well mixed throughout the culture).

3. Subsequent later migrations in history (e.g., the Tumulus culture or Celtic migrations) were more successful at moving people west than east, therefore the R-Z304 geographical locus shifted westward over time.

We cannot tell whether R-Z304 was in the Únětice culture or west of it, so in the maps displayed herein, we hedge our bets by placing them in the western part of the culture's extent.

7.4.11 R-Z156>Z306>Z307>Z304>BY12480

TMRCA: Assuming one SNPs since R-Z304 and a coverage of 16 Mbp, 2300 BC (95% c.i., 2631–2102 BC).

Ancient DNA: R-BY12482 has two ancient DNA results, R-FGC8365>DF96 has seven results, R-FGC8365>A10971 has three.

Modern testers: R-BY12480 contains R-DF96, which represents 37% of R-Z304.

Narrative: We can understand R-BY12480 best by considering R-FGC8395>DF96 separately, and focussing on its other sub-clades, R-BY12482 and R-FGC8395>A10971.

These are both considerably younger than R-BY12480, forming during the middle Bronze Age. They contain disproportionate amounts of ancient DNA, suggesting that they are (or were) larger than their data currently makes them appear. Unlike R-DF96 (or R-DF98), they are found in above expected numbers in south-eastern Europe (although there are still only three testers and two ancient DNA results from the House of Báthory to confirm this) and in Poland (again, only two families). Like R-DF96 (and R-DF98), they are strong in north-west Europe, but predominantly in Germany rather than France.

7.4.12 R-Z156>Z306>Z307>Z304>BY12480>BY12482

TMRCA: Family Tree DNA provides 1406 BC (95% c.i., 2065–850 BC).

Ancient DNA:

- I15950; 480–390 BC; middle aged La Tène Celt from north-west Bohemia.
- R58; 700–1500 AD; medieval Italian from Lazio (Italy); R-Y24836>BY12484.

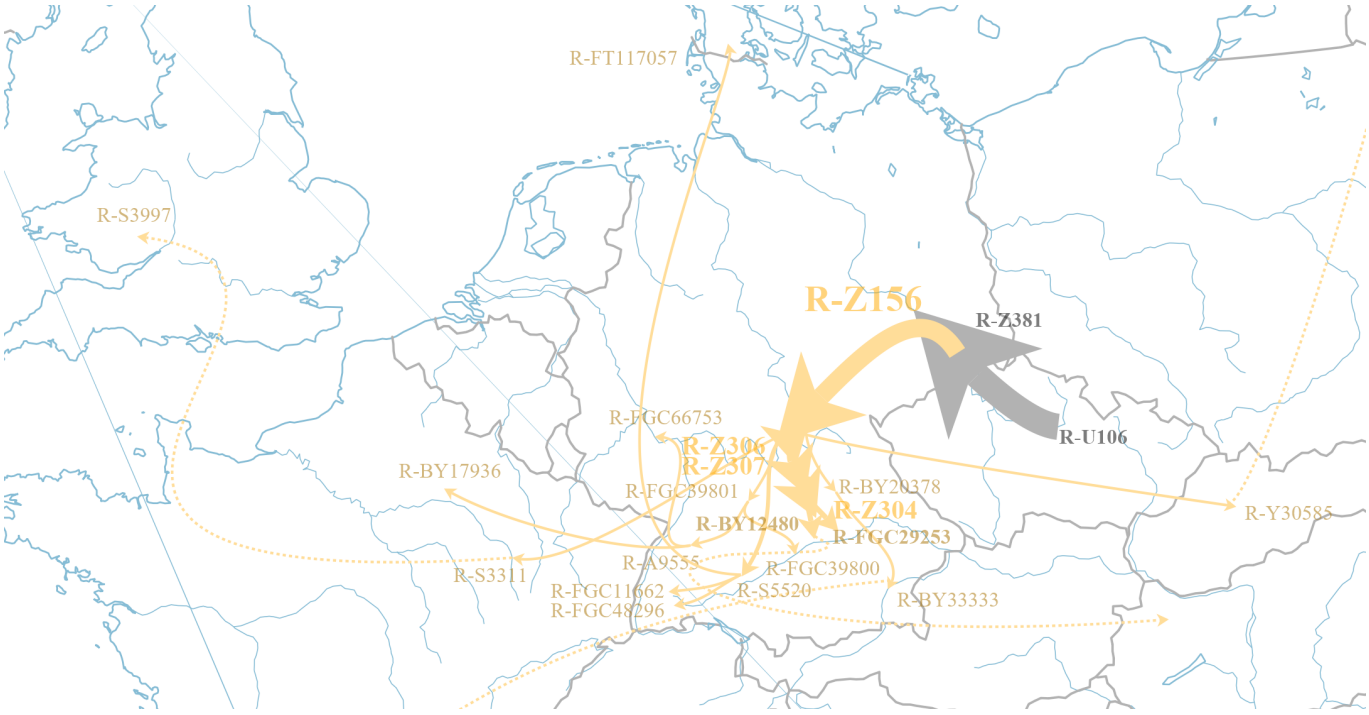

Figure 9: A best-guess map of the migrations of R-Z156 basal clades, based on their individual analysis. Dotted lines show smaller or recent migrations. This map is not expected to be entirely accurate.

Modern testers: 74 testers, 41 with European origins, 27 from the British Isles (24 of whom are English).

The 2000-year-old basal branch R-BY62658 contains three German testers, but most testers are within the 3200-year-old R-Y24836 branch. This splits into the Anglo-German (and one Croatian) R-Y65747 branch and the R-BY12484 branch. As well as containing the medieval Italian, R-BY12484 hosts the 2400-year-old German R-BY25572 and the 2000-year-old R-BY19410, which is mostly English but contains Polish and Macedonian testers.

Expansion: The expansion of R-BY12482 has been relatively constant thoughout history.

Narrative: Overall, the distribution of R-BY12482 has a locus that is farther west than the Únětice culture, but contains several groups to the east and south of its Czech homeland. The Celtic ancient DNA from Bohemia is also curious.

The presence of Germans and the Croatian at R-FT460054 and the Polish and Macedonian testers in R-BY19410 suggests that the English testers in these haplogroups migrated to England after their origin, therefore the migration more likely associated with the Anglo–Saxons (etc.) than the Romans.

7.4.13 R-Z156>Z306>Z307>Z304>BY12480>FGC8365>A10971

TMRCA: Assuming a coverage of 16 Mbp, R-FGC8365 has a TMRCA of 2231 BC (95% c.i., 2577–2014 BC). Family Tree DNA provides a TMRCA for R-A10971 of 1227 BC (1922–649 BC).

Ancient DNA:

- PCA0193; 1000–1200 AD; Greater Poland; R-BY18855>FTE67170.
- Two individuals from the House of Báthory, a Hungarian noble family descended from Briccius Báthory (d.∼1322). Part of the Gutkeled clan, who traditionally descend from Swabian brothers who emigrated from Baden-Württemberg.

Modern testers: 42 testers, 30 with European origins, 12 from the British Isles. The British Isles testers are dominated by the ∼900-year-old Irish R-BY166050 Smith family.

Geographically (also taking into account ancient DNA), the haplogroup appears to split into an older group that focusses more towards Germany and Poland, with offshoots to Fenno-Scandia, and the younger (2000-year-old) R-BY18860 group, which contains southern German, Hungarian and Bosnian testers.

Narrative: The geographical wanderings of this haplogroup are a little unclear, but they appear to have several migrations to both the north-east and south-east of the main R-Z304 and R-BY12840 loci, thus represent a more eastern offshoot compared to R-DF96's westward migration.

7.4.14 R-Z156 minor near-basal clades: conclusions

Figure 9 shows a summary of the above discussion. Here, some arbitrary choices have had to be made about the precise direction and timing of R-Z156's migrations, which are far from clear. A lot of migration has gone on in this ancient haplogroup and it is difficult to pin down timings and directions precisely.

Combining the southern distribution of modern R-Z156 testers today with the general lack of testers east of the Czech Republic, and it seems less likely that R-Z156 simply stayed in Bohemia between R-U106 and R-Z304, and more likely that it detoured west. The precise location is unclear, but southern Germany is a sensible origin for most of R-Z156's major sub-clades, and would give an easy route back into Bohemia for I7196.

The presence of R-Z156 (specifically, R-Z304) in a region spanning from Holland to Bohemia by ∼1300 BC indicates significant migration during the first 1000 years. The Tumulus culture may play a role here, being the dominant culture in the preceding few centuries. From there, many R-Z156 would have found themselves in the Urnfield culture and later Celtic and Germanic cultures, as evidenced by the abundance of ancient DNA from both of these latter groups in central and southern Europe.

We also see suggestions of early migrations into France, which may have equally avoided the Únětice culture altogether or arrived in a post-Únětice migration. We also see scattered lineages travelling east. Some of these trends are also present in the major sub-clades R-DF98 and R-DF96 (see below) which could share further light on R-Z156's early expansion.

7.5 R-U106≫Z381>Z156>Z306>Z307>Z304>FGC29253>DF98

To be added

7.6 R-U106≫Z381>Z156>Z306>Z307>Z304>BY12480>FGC8365>DF96

To be added

7.7 R-U106>Z2265>BY30097>Z381>S1688

To be added

7.8 R-U106>Z2265>BY30097>Z381>S1688

To be added

7.9 R-U106>Z2265>BY30097>Z381>L48

To be added

8 Conclusions

8.1 Phylogeography

- Phylogeography still lacks a mathematical model that can robustly define a haplogroup's origin. Problems include accounting for biased testing among populations, the asymmetric migration of testers, inaccuracies in genealogical information, and the incompleteness and inaccuracy of associated geographical information.
- Strong variation in genetic testing exists between different countries. This is most problematic where well-tested and poorly-tested countries border each other. The gap between the best-tested country (Scotland) and the worst-tested countries (France, Russia) is a factor of approximately 38.
- The typical genealogist taking a genetic test knows their ancestry to somewhere between 1700 and 1850 AD.
- Paternal ancestry information can be given as country of origin, detailed earliest-known ancestor information, and latitude/longitude. Where these can be compared, there are reasonable error rates between them, reaching as high as ∼67% errors in country of origin for those attesting an (unspecific) origin in the UK but have stated a paternal ancestry elsewhere (normally in the USA). The average rate across all testers is ∼41% error.
- This will likely significantly over-estimate the true error rate, as many American families will have correctly surmised their origin is in their stated parts of Europe, but they simply lack the genealogical evidence to prove that conclusively. However, "guestimated" error rates of ∼10–20% mean that country flags still provide concerningly noisy data, and should be verified against statements of genealogy where practical.
- NPEs and drift between socially recorded genealogies and genetic inheritance of data are a concern, but only contribute meaningfully when pedigrees are longer than average and have not yet been triangulated.

8.2 The initial spread of R-L151

- The importance of the ancient DNA burial PNL001 is asserted. This earliest R-U106 burial, right at the start of the Corded Ware Culture migration, effective enforces the origin of R-L151 to pre-date the CWC migration. This places the origin of R-L151 among the unidentified pre-CWC-migration population, hypothesised to be either in the forest belt around Latvia and Muscovy, or in the Yamanya-derived populations of the central Danube.
- The need for a split of R-ZZ11 within the Corded Ware Culture or its derivatives, and the lack of R-L151 basal haplogroups that derive from east or south-east of the CWC is used to limit the oldest likely age of R-L151.
- This provides a TMRCA of 3115 BC (3222–3029 BC, 68% c.i.; 3366–2972 BC, 95% c.i.; 3507–2937 BC, 99.5% c.i.).
- The average growth rate during the first ∼625 years of R-L151 was at least as high as the growth rate in humans today (≳1% per annum). This rate does not account for branches that have died/daughtered out, so may be a significant under-estimate. Individual families must have had much higher rates of growth.

8.3 The initial spread of R-U106

- The presence of PNL001 during the earliest parts of the CWC suggests Bohemia is (or is close to) the location in the CWC from which R-U106 spread.
- The lack of a strong and diverse R-U106 presence in the Czech Republic following PNL001 suggests that most R-U106 branches did not stay in Bohemia, though R-Z156 is consistently found.
- Minor R-U106 clades have been found in Spain during the Bell Beaker period, indicating that R-U106 indeed played a not-particularly-successful rôle in the Bell Beaker resurgence into Iberia.
- R-Z156 is common in pre-Germanic burials across the southern front of R-U106's expansion, suggesting that R-U106xZ156 haplogroups were restricted to the north of this frontier.
- R-Z18 is a very strong component of burials in Denmark and southern Sweden from the Bell Beaker period (∼2300 BC) onwards. Although a few R-U106xZ18 burials are also seen, their general lack suggests that R-U106xZ18 haplogroups were restricted to south of this frontier.
- R-Z301 remains surprisingly absent in burials before 300 BC. Its hypothesised origin (based on locations where ancient DNA testing is absent yet R-Z301 frequent in modern populations) is either in or close to northern or western Germany. Ancient DNA only attests that R-Z301 "broke out" of this region around 300 BC to the north and during the Migration Age to the south.
- The expansion of the minor clades of R-U106 is most consistent with a generalised push into the Corded Ware Culture regions of northern Germany during the first few centuries of R-U106's European growth. However, R-U106 as a whole appears to have been a relatively homogeneous unit during this period. This potentially gives it a role in the Single Grave Culture, within the later phases of the Corded Ware cultural umbrella.
- Following this period, several R-U106 sub-clades become entrenched in the Bell Beaker movement. Evidence for Bell Beakers among ancient DNA include R-Z18 (CGG107465) in Denmark, and R-FGC396 (CGG_2_023808) and R-S18632 (CGG_2_023745) in Spain. R-Z301>FGC13959 may also be present in Denmark (CGG106838), but this is less certain due to conflicting TMRCAs and ^{14}C dates, hence the possibility for misassignment or contamination.
- The R-Z18 Bell Beaker component quickly morphed into the early Nordic Bronze Age groups, from which all or almost all of modern R-Z18 testers appear to descend. A large portion of these lines appear to be associated with the early Germanic groups that spread both north and south during the last millennium BC.

8.4 Recommendations

- The development of phylogeographical models that better account for known issues in the input data.
- Improvement of the input data at a user level, by encouraging genealogists to update and make more accurate their paternal ancestry information.
- Errors in provided country-level origins mean that phylogeographical analyses need to take into account information from publicly available genealogies, such as the earliest-known ancestor information and quoted latitude/longitude of origin.
- Triangulation of long genealogies is important to remove uncertainty in both those genealogies and the origins of the associated haplogroup.

- Engagement with the ancient DNA community to encourage better genetic recovery of individual samples, particularly on the Y chromosome.
- Development of a robust methodology and accepted practice to identify *de novo* mutations in ancient DNA Y chromosomes, in order to allow TMRCAs from ancient DNA.

A Sources of historical census information

Table 5 lists the sources of historical populations used on a per-country basis. Additional values were sourced from the *"Demographics of (country)"* pages from Wikipedia, and the *"Total population (country)"* pages on `http://www.statista.com`. Values for Albania, Romania, Cyprus and Malta are taken from `http://www.familysearch.org/wiki/en/`.

Table 5: Sources of historical census information. Dates rounded to closest 10 years

Country	Date range	Source
Various	1700, 1820	[26], Table B.10
UK	1800–1850	UK census[a]
England	1700	https://en.wikipedia.org/wiki/Demography_of_England
Scotland	1710	https://www.gov.scot/publications/scotland-future-opportunities-challenges-scotlands-changing-population/pages/4/#
Scotland	1760	https://www.nrscotland.gov.uk/research/guides/census-records/webster%E2%80%99s-census-of-1755
Ireland	Various	[27]
Ireland	1810	https://www.libraryireland.com/articles/ModIrelandDecreasePopulation/index.php
N. Ireland	1821–1851	https://www.statista.com/statistics/1015418/population-northern-ireland-1821-2021/
Ireland	1700–1710	https://www.libraryireland.com/articles/populationirelandDPJ/index.php[a]
Isle of Man	1821–1851	https://www.gov.im/media/207874/2001censusreportvolume2.pdf
Poland	pre-1800	http://www.cicred.org/Eng/Publications/pdf/c-c43.pdf, corrected by area
Czechia	Various	https://csu.gov.cz/produkty/czech-demographic-handbook-2022
Austria	Various	https://www.statistik.at/en/statistics/population-and-society/population/population-stock/population-by-age-/sex
Denmark	Various	https://web.archive.org/web/20170908201150/http://statistikbanken.dk/statbank5a/default.asp?w=1024
Denmark	Various	[28]
Norway	Various	[28]
Russia	1820	[28]
Spain	1830	[28]
Portugal	1841	[28]
Sweden	Various	http://runeberg.org/sverig01/0106.html
Finland	Various	http://tilastokeskus.fi/til/vrm_en.html
Iceland	Various	http://px.hagstofa.is/pxen/pxweb/en/Ibuar/Ibuar__mannfjoldi__1_yfirlit__yfirlit_mannfjolda/MAN00000.px
Greenland	Various	[29]
Russia	Various	https://dmorgan.web.wesleyan.edu/materials/population.htm
Belarus	—	Estimated as 5.3% of the Russian population
Moldova	Various	https://www.researchgate.net/publication/330411227_HISTORICAL_MOLDAVIA-_FROM_DEMOGRAPHIC_EXPANSION_TO_A_SHRINKING_REGION[b]
Bosnia	Various	https://clio-infra.eu/Countries/BosniaandHerzegovina.html
Serbia	Various	https://www.rastko.rs/istorija/srbi-balkan/sradovanovic-demography.html
Spain	1812	https://www.napoleon-series.org/research/miscellaneous/c_population.html
World	Various	https://en.wikipedia.org/wiki/Estimates_of_historical_world_population

[a]Portions in Republic of / Northern Ireland apportioned using the 1821 ratio of 1380:5422. [b]Estimated as 30% of historic Moldavia numbers.

B Glossary

Haplogroup terminology

See the list of genetics terms for symantic definitions of haplogroup, clade, etc., as applied to this document. Haplogroups can either be referred to by their long (e.g., R1b1a1b1a1a1) or short (e.g., R-U106) format. The former uses a tree-based structure whereby R branches into R1 and R2, R1 branches into R1a and R1b, etc. The latter bases its name on the major haplogroup (R) and one of the SNPs that defines the branch (U106). The long format is now generally disused as it can change with every tree revision, though the top groups like R1a and R1b are often still seen, as these are fairly stable. The short format is more stable, but still subject to change if a haplogroup is split in half by the formation of a new branch.

In this document, we will refer to haplogroups like R-U106, meaning the group of men who are U106+. However, we also use:

- R-U106* for men who are U106+ but are tested negative for known downstream haplogroups (in this case, R-Z2265 and R-A2150);
- R-U106xZ2265 for someone who is tested U106+ but Z2265- (therefore could be R-A2150 or R-U106*);
- R-U106?Z2265 for someone who is tested U106+ and is possibly Z2265+ based on poor-quality data.

This latter case is common in ancient DNA, where we can also find the use of, e.g., "R-U106 (pre-A2150)", meaning an ancient DNA sample that is U106+ and positive for some of the SNPs in that define the R-A2150 haplogroup, but negative for some others.

Geographical terms

The regions **North-West Europe**, **North-Central Europe**, **Eastern Europe**, **South-Eastern Europe** and **Mediterranean** are used as defined in Table 1. The term **Europe** includes all of Russia, on the understanding that the majority of the Russian population lives in (or is patrilineally descended from) the part of Russia within the European continental mass (i.e., west of the Urals). Turkey is included in Europe for the purposes of this paper, but the Caucasus nations are not.

- **British Isles:** for the purposes of this document, the islands of Great Britain, Ireland and other islands in the archepeligo, as distinct from the countries of the Ireland, the UK, and its constituent nations.
- **Continental Europe:** the main landmass of Europe, excluding the British Isles, Iceland, the Faeroes and Fennoscandia.
- **Fennoscandia:** the peninsula containing Norway and Sweden, plus Finland. Nominally including the Kola peninsula and Karelia, though these are not included here.
- **Nordic countries:** including Norway, Sweden, Denmark, Finland, Iceland, the Faeroes and Greenland.
- **Scandinavia:** the peninsula of Norway and Sweden, plus Denmark. Distinct from the Nordic Countries.

List of acronyms and abbreviations

- **c.i.:** confidence interval. For example, a 95% confidence interval means there is formally a 95% probability that the true value lies within the stated range (though this normally does not count for "unknown unknowns" — for example, in the ancient DNA analysis, this does not account for the caveats mentioned in Section 5.1).
- **CWC:** Corded Ware Culture. The dominant culture in northern Europe during the period ∼3000–2300 BC.
- **EKA:** earliest known ancestor. For Y-DNA, this represents the oldest person known on a person's male line. Also called a most-distant known ancestor (MDKA).
- **GD:** A count of the number of mutations that separate two individuals. While this can be applied to counts of Y-SNPs, it normally refers to Y-STRs. Most commonly (e.g., as done by Family Tree DNA), the *inifinite-alleles method* is used, which counts the number of mis-matching Y-STRs. The *stepwise method* is sometimes also used, which sums the difference in the number of repeats (e.g., testers with a Y-STR with alleles 13 and 15 would have a genetic distance of two).
- **ISOGG:** International Society of Genetic Genealogists.
- **mt-DNA:** DNA stored on in the mitochondria, passed down from mother to daughter.
- **MNP:** multi-nucleutide polymorphism. This is where several base pairs change together, e.g. ACGT → TGCA.

- **MRCA:** most recent common ancestor. For Y-DNA, this represents the last person in a male-line family tree that is related to two different Y-DNA testers, with each tester then descending from different brothers. For example, my cousin's paternal grandfather is my paternal grandfather, so that grandfather is our MRCA.
- **MPE:** misattributed-parentage event. Any event where the genetic parent of the child is not that listed in a genealogy. This includes NPEs, but also issues arising from poor genealogy (either through lack of rigour or lack of records).
- **NPE:** Strictly, a non-paternity event or, more loosely, "not the parent expected". Strict definitions are that the genetic father of a child is not the father named on a birth certificate (or similar document). Loose definitions are any case where the EKA is not what is stated by the genetic tester, including MPEs and SDEs, which are more testable in phylogenetic studies.
- **SDE:** surname discontinuity event. Any event that changes the surname of the paternal line. This includes NPEs, some MPEs, adoptions, intentional surname changes, etc.
- **T2T:** Telomere-to-Telomere Consortium. This group has been using long-read technologies to completely sequence the human genome, including the harder-to-read parts of the Y chromosome.
- **TMRCA:** Time to most-recent common ancestor. Literally, the number of years since the common ancestor of two testers lived. For consistency, it is best calculated from the birth date of the modern testers (usually a date between 1950 and 1960 AD is assumed) and results in an estimate of the number of years since their MRCA was born. The term TMRCA is also be used (slightly erroneously) to refer to the corresponding calendar date.
- **Y-DNA:** DNA stored on the Y chromosome, passed down from father to son.
- **Y-SNP:** Single nucleotide polymorphisms on the Y chromosome. These represent the change of a single base pair. The fewer mismatching Y-SNPs two testers have in a sequence of DNA, the closer they are related.
- **Y-STR:** Short tandem repeats on the Y chromosome. These represent short, repeating structures within the genome. Changes in the number of repeats occur over time. The differences in the number of repeats in a set of Y-STRs defines a genetic distance (GD), which is linked to how far in the past two men are related.

Genetics terms

- **Ancient DNA:** DNA recovered from ancient remains, normally remains discovered via archaeological digs. Ancient DNA is often heavily degraded, often leading to less certainty (lower coverage) of critical Y-SNPs, hence a greater number of false-positive SNPs. This often means that only recovery of an approximate haplogroup is possible from ancient DNA, with no novel SNPs being identified.
- **Allele:** The genetic result read for a particular genetic marker. For Y-SNP tests (including next-generation seqeuencing tests), this is normally the base pair read at the SNP's position on the reference genome (or [plural] base pairs read for an MNP, insertion or deletion). For Y-STR tests, this is the number of repeats recorded, or the Y-STR's "value".
- **Base pair:** Pairs of four possible molecules (chemical bases), which encode genetic data in DNA. These can be adenine, thymine, guanine or cytosine, abbreviated as A, T, G and C (RNA also includes the base uracil, abbreviated U). Specific bases always bond to other bases to form base pairs (A always to T, G always to C). These base pairs bond in sequence to the phosphate backbone to form the DNA double helix. The pairing means that the second strand in the DNA double helix can always be inferred in only the first is read.
- **Clade:** The group of individuals descending from a common ancestor in whom a novel genetic marker formed. Often used interchangably with *haplogroup*, despite slight differences in meaning.
- **Deletion:** The removal of some genetic data from a person's genome, e.g., AATGCC $\rightarrow$ AACC.
- **Family Tree DNA:** A genetic testing company in Houston, Texas, which specialises in the recovery of Y-DNA and mt-DNA from living testers.
- **Founder effect:** The very successful foundation of a new haplogroup by a man, which leads to an outsize sub-clade within a haplogroup. If this success was associated with a migration, it may move the geographical locus of the entire parental haplogroup away from its location of origin.
- **Genealogies:** A list of historical paper records — a "paper trail" — linking genetically tested individuals to their ancient ancestors.
- **Haplogroup:** A group of genetic tests that share common sets of genetic markers (haplotypes). Often used interchangably with *clade*, despite slight differences in meaning.

- **Haplotree:** The name given to the phylogenetic tree by Family Tree DNA, after the practice of putting haplogroups onto a family tree structure.
- **Haplotype:** A set of genetic results.
- **Insertion:** The addition of some genetic data into a person's genome, e.g., AACC → AATGCC.
- **Man:** The term "man" in the context of this document represents a genetic man in the strictest sense, i.e., one who has had at least one copy of the SRY gene. This is without prejudice to genetically intersex individuals, who may host incomplete or multiple Y chromosomes, individuals whose social gender does not match their genetic sex, or genetic men with mosaic loss of the Y-chromosome.
- **Male-line ancestry:** The male-only part of the family tree, i.e., a person's father, their father's father, their father's father's father, etc.
- **Most-recent known haplogroup:** The youngest named haplogroup to which a man belongs. For example, in Family Tree DNA's Discover tool, on the Ancestral Path tab, this is the last haplogroup listed in the tree, and this is the haplogroup that shows up in user's accounts. Family Tree DNA refers to this as a "terminal" haplogroup. However, it is important to note that users may belong to a more recent haplogroup that is not yet known, and the discovery of such a haplogroup would mean their "terminal" haplogroup is updated.
- **Next-generation sequencing:** Tests available to consumers since 2014, which have allowed large portions of the genome to be extracted in their entirety, allowing detection of novel Y-SNPs and recovery of more Y-STRs. The most common next-generation seqeuencing test is Family Tree DNA's Big Y test.
- **Paper trail:** A genealogical record that links different individuals of the same family together.
- **Phylogenetic tree:** A tree of relationships built up from interpretation of genetic data. Also known as a haplotree.
- **Phylogeography:** The study of how genetic groups have expanded and diversified over the course of history.
- **Terminal haplogroup:** see most-recent known haplogroup.
- **Triangulation:** The process of obtaining two tests from two descendants of a person to check the accuracy of genealogy. If the tests match, their genealogies can be considered proven back to their most-recent common ancestor (see MRCA).
- **X chromosome:** One of the two sex chromosomes in mammals, the other being the **Y chromosome**. The X chromosome is inherited with only one copy by genetic men and two copies by genetic women.
- **Y chromosome:** One of the two sex chromosomes in mammals, the other being the **X chromosome**. The Y chromosome is identified by the SRY gene, which defines genetic maleness among mammals. The Y chromosome is passed strictly from father to son.

References

1. Sims, L. M., Garvey, D., and Ballantyne, J. *Human Mutation* **28**(1), 97–97 (2007).
2. Sahakyan, H., Margaryan, A., Saag, L., Karmin, M., Flores, R., Haber, M., Kushniarevich, A., Khachatryan, Z., Bahmanimehr, A., Parik, J., Karafet, T., Yunusbayev, B., Reisberg, T., Solnik, A., Metspalu, E., Hovhannisyan, A., Khusnutdinova, E. K., Behar, D. M., Metspalu, M., Yepiskoposyan, L., Rootsi, S., and Villems, R. *Scientific reports* **11**(1), 6659–6659 (2021).
3. Fort, J. *Neolithic Transitions: Diffusion of People or Diffusion of Culture?*, 327–346. Springer International Publishing, Cham (2023).
4. Busby, G. B. J., Brisighelli, F., Sánchez-Diz, P., Ramos-Luis, E., Martinez-Cadenas, C., Thomas, M. G., Bradley, D. G., Gusmão, L., Winney, B., Bodmer, W., Vennemann, M., Coia, V., Scarnicci, F., Tofanelli, S., Vona, G., Ploski, R., Vecchiotti, C., Zemunik, T., Rudan, I., Karachanak, S., Toncheva, D., Anagnostou, P., Ferri, G., Rapone, C., Hervig, T., Moen, T., Wilson, J. F., and Capelli, C. **279**(1730), 884–92 08 (2011).
5. Larmuseau, M., Vanoverbeke, J., Van Geystelen, A., Defraene, G., Vanderheyden, N., Matthys, K., Wenseleers, T., and Decorte, R. *Proceedings of the Royal Society B: Biological Sciences* **280**(1772), 20132400 (2013).
6. Dahlén, T., Zhao, J., Magnusson, P. K., Pawitan, Y., Lavröd, J., and Edgren, G. *Journal of Internal Medicine* **291**(1), 95–100 (2022).

7. McColl, H., Kroonen, G., Moreno-Mayar, J. V., Valeur Seersholm, F., Scorrano, G., Pinotti, T., Vimala, T., Sindbæk, S. M., Ethelberg, P., Fyfe, R., Gaillard, M.-J., Larsen, H. M. E., Mortensen, M. F., Demeter, F., Jørkov, M. L. S., Bergerbrant, S., Damgaard, P. d. B., Allentoft, M. E., Vinner, L., and et al. *bioRxiv* (2024).

8. Papac, L., Ernée, M., Dobeš, M., Langová, M., Rohrlach, A. B., Aron, F., Neumann, G. U., Spyrou, M. A., Rohland, N., Velemínský, P., Kuna, M., Brzobohatá, H., Culleton, B., Daněček, D., Danielisová, A., Dobisíková, M., Hložek, J., Kennett, D. J., Klementová, J., Kostka, M., Krištuf, P., Kuchařík, M., Hlavová, J. K., Limburský, P., Malyková, D., Mattiello, L., Pecinovská, M., Petriščáková, K., Průchová, E., Stránská, P., Smejtek, L., Špaček, J., Šumberová, R., Švejcar, O., Trefný, M., Vávra, M., Kolář, J., Heyd, V., Krause, J., Pinhasi, R., Reich, D., Schiffels, S., and Haak, W. *Science Advances* **7**(35), eabi6941 (2021).

9. Vokolek, V. *Archeologické Rozhledy, XXXIII* , 481–485 (1981).

10. Ryan-Despraz, J. *Practice and prestige: An exploration of neolithic warfare, Bell beaker archery, and social stratification from an anthropological perspective*. Archaeopress Publishing, (2022).

11. Gimbutas, M. *The prehistory of eastern Europe*. Number 20. Peabody Museum, (1956).

12. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S. L., Risch, R., Rojo Guerra, M. A., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K. W., and Reich, D. *Nature* **522**(7555), 207–211 June (2015).

13. Nordqvist, K. and Heyd, V. In *Proceedings of the Prehistoric Society*, volume 86, 65–93. Cambridge University Press, (2020).

14. Furholt, M. In *Proceedings of the Prehistoric Society*, volume 80, 67–86. Cambridge University Press, (2014).

15. Linderholm, A., Kılınç, G. M., Szczepanek, A., Włodarczak, P., Jarosz, P., Belka, Z., Dopieralska, J., Werens, K., Górski, J., Mazurek, M., et al. *Scientific Reports* **10**(1), 6885 (2020).

16. Kershaw, J. and Røyrvik, E. C. *Antiquity* **90**(354), 1670–1680 (2016).

17. Helgason, A., Einarsson, A. W., Guðmundsdóttir, V. B., Sigurðsson, Á., Gunnarsdóttir, E. D., Jagadeesan, A., Ebenesersdóttir, S. S., Kong, A., and Stefánsson, K. *Nature genetics* **47**(5), ng–3171 (2015).

18. McDonald, I. *Genes* **12**(6), 862 (2021).

19. Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., et al. *Nature* **555**(7695), 190–196 (2018).

20. Patterson, N., Isakov, M., Booth, T., Büster, L., Fischer, C.-E., Olalde, I., Ringbauer, H., Akbari, A., Cheronet, O., and et al. *Nature* **601**(7894), 588–594 January (2022).

21. Brittain, M. (2017).

22. Martiniano, R., Caffell, A., Holst, M., Hunter-Mann, K., Montgomery, J., Müldner, G., McLaughlin, R. L., Teasdale, M. D., Van Rheenen, W., Veldink, J. H., et al. *Nature communications* **7**(1), 10326 (2016).

23. Myres, N. M., Ekins, J. E., Lin, A. A., Cavalli-Sforza, L. L., Woodward, S. R., and Underhill, P. A. *Croatian Medical Journal* **48**(4), 450 (2007).

24. Yediay, F. E., Kroonen, G., Sabatini, S., Frei, K. M., Frank, A. B., Pinotti, T., Wigman, A., Thorsø, R., Vimala, T., McColl, H., Moutafi, I., Altinkaya, I., Ramsøe, A., Gaunitz, C., Renaud, G., and et al. *bioRxiv* (2024).

25. Gagnon, D. and Beddows, P. *Foundations* **15**, 64–85 (2022).

26. Maddison, A. *The World Economy*. (2006).

27. Grada, C. O. *Annales de démographie historique* , 281–299 (1979).

28. Drake, M. *Scandinavian Economic History Review* **13**(2), 97–142 (1965).

29. Marquardt, O. *Études/Inuit/Studies* **26**(2), 47–69 (2002).