U106 explored: its relationships, geography and history

The 2015 report to the U106 group (September update)

Principal investigator: Iain McDonald

Contents

Methodology & Background

WARNING!

Everything presented in this document is wrong. The question is: how wrong? It is my hope that there is enough in here that is only a *little bit* wrong for it to be useful.

The dates I compute here were outdated before I finished compiling the information. More tests have resolved more branches of our family tree. Further archaeological DNA was published that altered the subtleties of the migrations portrayed at the end of this document. This is a very active field of research, advancing at an impressive rate. So please take everything in this document with a hefty pinch of salt and note of skepticism.

FOREWORD

The decisions of our ancestors have shaped the world we live in today, whether that decision was to fight or flee, what to hunt, who one should marry, or simply what to get for breakfast. Those countless decisions became manifest in their successes or failures in life, whether they survived to produce a family, and ultimately the fact that some of those families prospered and eventually produced you and I. In this work, we attempt to re-trace some of those decisions back in time to find who our ancestors were, and what decisions were made that allowed them to play a critical role in the history of Europe and the wider world.

Our particular story here concerns a family whose descendants carry a particular genetic mutation - worn like a molecular badge - which allows us to identify them as sharing a single common ancestor in which this mutation first arose. That mutation is named U106, or alternatively S21, and is the result of a simple typographical error that happened around 4500 years ago, where one encoding molecule was misread among the 59 million that made up the Y-chromosome of a particular cell. That cell grew into a man, and that man is the 125-times great-grandfather (or thereabouts) of about one in eight men of European descent today.

This document attentions to trace his descendants and the paths they took throughout history, and maps their distribution throughout Europe close to the present day.

METHODS: DNA TESTING BASICS

This report is an analysis of the Y chromosome, which is passed from father to son. It is therefore only a study of male lines: a person's father's, father's, father's, … father. It can therefore be used to trace the history of a surname, and uncover "superfamilies" which were founded before the age of surnames.

DNA is made up of four bases: A, C, G and T, and can be read out as a string of these letters. A single person's DNA means nothing. Genetic genealogy relies on a comparison between two or more people's DNA. The differences between them identify mutations that have happened in the transfer of the genetic code from parent to child. These mutations can be used to work out relationships, and the time since their most-recent common ancestor (TMRCA).

METHODS: Y-STR TESTING

There are two different kinds of DNA that are used to determine people's relationships to each other. The most commonly taken is an STR (Short Tandem Repeat) test, advertised at Family Tree DNA as a series of 12, 25, 37, 67 or 111 markers. These markers take the form of a short section of DNA that repeats a certain number of times. Mutations can cause this number to increase or decrease. A hypothetical example would be:

`DYS1234 = 4 TACATACATACATACA`

which could mutate to:

`DYS1234 = 5 TACATACATACATACATACA`

by gaining a repeat.

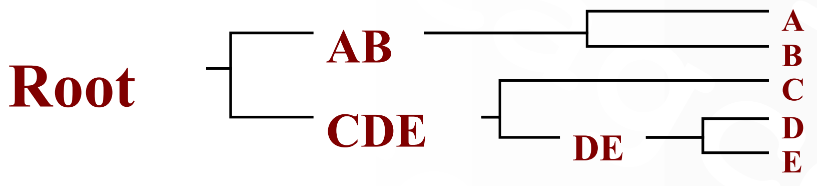
If most people have DYS1234=4 and some people have DYS1234=5, we presume that "4" is the ancestral value and that the people with "5" are more closely related.

Things are rarely that simple, as the same mutation can happen in different branches, STR markers can mutate back to their ancestral values, and a lot of poorly understood factors make them prefer certain values over others. For these reasons, they stop being very accurate tools on long timescales, and are not absolutely foolproof for creating these family groups. We tend to need two or more shared mutations to ensure a person belongs to a specific group.

Using a series of these mutations, we can build a relationship tree for families, e.g., for:

DYS	393	390	19	391	385	426	388	439	389i	392	389ii
A:	13	24	14	11	11-15	12	12	12	12	13	29
B:	13	24	14	10	11-15	12	12	12	13	13	29
C:	13	24	14	11	11-14	12	12	13	13	13	29
D:	13	23	14	11	11-14	12	12	13	13	13	29
E:	13	23	14	11	11-14	12	12	13	13	13	29

we presume the group CDE are more closely related because of the DYS439=13 mutation, with DYS390=23 defines are group within this (DE). DYS385=11-15 defines another group (AB). Thus:



METHODS: Y-SNP TESTING

The second test we perform is Y-SNP testing. Outside of the repeating STR regions, DNA is more of a genetic jumble. As material is passed down, parts of the code can be inserted:

`ATGCTGATCGC → ATGCTGATAGATCGC` ,

deleted:

`ATGCTGATAGATCGC → ATGCTGATCGC` ,

or mutated:

`ATGCAGATCGC → ATGCTGATCGC` .

Sites of these latter mutations are known as single nucleotide polymorphisms, or SNPs. These SNPs are very reliably passed on from father to son, so they can clearly identify a family branch without the ambiguity than STRs provide.

SNPs can be tested individually through Sanger sequencing, as used conventionally by Family Tree DNA and YSeq. They can also be tested *en masse* and new SNPs discovered through 'second-generation' tests such as the Illumina dye sequencing used in Family Tree DNA's BigY or Full Genome Company's Y Elite and Y Prime.

We use these SNP tests to create the backbones structure of the human Y-DNA tree, draping over it the STR results of all testers to flesh out the branches. For a full understanding of the human male-line family tree, we require comprehensive SNP testing of every branch, backed by STR results to compare with the larger STR databases.

FUNCTION OF THE U106 GROUP

The U106 group facilitates these comparisons by providing a place where individual testers can share their data, regardless of the company and country of origin. The group provides expertise to analyse that data, and can make recommendations for people to get the greatest return from each test. By collecting this data together, we provide a sample size greater than almost every professional study (even though it is not so homogeneously sampled as such studies).

Although U106 encompases a lot of people, perhaps 3% of human male lines, it is a comparatively small twig of the human Y-DNA tree. By focussing on this single twig, we can provide a greater depth of analysis and understanding than broader-ranging professional scientific studies are able to, and drill deeply into the recent history of individual families.

This approach relies on the generosity of individuals who are willing to share the details of their genome with the community. In return, they get to learn more about their family history. This work would not have been possible without them. Nor would it have been possible without the support of the rest of the U106 team - primarily Charles Moore and Raymond Wing, who work tirelessly behind the scenes to keep the ship afloat and on the right heading, and David Carlisle and Andrew Booth for sorting the details of BigY. Kudos also goes to Dr. Tim Janzen and Prof. Ken Nordtvelt for detailed help with different aspects of STR age analysis. Finally, thanks to the innumerable members of the U106 Yahoo forum who have contributed in many different ways to the success of this project.

NEXT-GENERATION TESTING

“Next-generation” tests like Family Tree DNA’s BigY and Full Genome Company’s Y-Prime and Y-Elite products offer an unparalleled chance to uncover new SNP mutations, insertions and deletions (indels) in your DNA. These are the only reliable tool we have for determining new structures within the Y-DNA tree (clades) and the only accurate way of obtaining dates we have. Of the 59 million base pairs in the human Y chromosome, BigY and Y-Prime test a little over 10 million, and Y-Elite tests around 16 million.

CLADE IDENTIFICATION

Clade identification typically progresses as follows. When a new test arrives, we get two sets of summary data: the coverage of the test, and the differences that test has from a known sequence. These are reported as positions along the chromosome, e.g.:

```
     chrY   2660548   2665410
```

means the test covers all base pairs between these two positions, and:

```
     chrY  2661694    .    A    G    1484.13   PASS .    GT   1
```

means that a mutation from A to G has occurred at position 2661694. In this case, this SNP has been later given a name, L311, which typically replaces this number. Occasionally, SNPs may be rejected if they have a low quality score:

```
     chrY  2649856    .    G    .    150.356   REJECTED   .    GT   0
```

SNPs from each test are compared to each other, e.g.:

```
7246726  7246726           7246726  7246726  7246726  7246726
23612197          23612197 23612197
                                    19047132 19047132
6788390  6788390
                  22178569 22178569
13494176                   13494176
22191144
         7906217
                  22758149
                                    17735808
                                             23165645
                                                      19035709
                                                             14991735
```

Names of known SNPs are filled in and singletons are put together:

```
Z381   Z381         Z381   Z381   Z381    Z381
L48           L48   L48
                           Z307   Z307
Z9     Z9
              L47   L47
Z2001               Z2001
SINGLETONS:
Z8     S6909  Z159         DF96   S1911   S5520  Z18
```

Comparisons to the coverage file let us fill in “no calls” (nc) and most false negatives (+?):

```
Z381   Z381   (+?)  Z381   Z381   Z381    Z381
L48    nc     L48   L48
                           Z307   Z307
Z9     Z9
              L47   L47
Z2001               Z2001
SINGLETONS:
Z8     S6909  Z159         DF96   S1911   S5520  Z18
```

Inconsistent SNPs (Z2001) are treated as false positives and removed:

```
Z381   Z381   (+?)  Z381   Z381   Z381    Z381
L48    nc     L48   L48
                           Z307   Z307
Z9     Z9
              L47   L47
SINGLETONS:
Z8     S6909  Z159         DF96   S1911   S5520  Z18
```

Providing a tree like that produced for U106 by Andrew Booth.

CHARACTERISATION OF FTDNA BIGY

We now have sufficient tests that we can perform a fairly rigorous characterisation of BigY. The analysis presented below is based on 408 BigY tests: the entire analysed sample as of 27 May 2015.

The average BigY test comprises of 10,585,146 base pairs (standard deviation 311,007) over 11,593 regions (st.dev. 2511). Typically 130 SNPs are called in each file including 14 novel variants (new SNPs private to this test).

A problematic region exists around position 22,400,000. Many SNPs are correctly called in this region, but there are a lot of falsely called SNPs too. This region of the Y chromosome is very similar to one on the X chromosome, and coverage of this region is very low. Many larger indels in this region are falsely reported as a series of SNPs. These often show up as singletons and confound later dating operations. For many applications, include dating of SNP ages, I have removed the entire DYZ19 region between positions 22216800 and 22512940 and do not use any SNPs found here. Typically 102,700 base pairs are called in this region.

Other problematic regions exist, but they are less significant, and do not greatly affect the overall results presented here.

From the first 319 BigY tests, the typical overlap between two tests (excluding DYZ19) is 10,176,279 base pairs (st.dev. 267,358), or 97.1% overlap. For two given tests, roughly 2.9% of SNPs will not be called in the matching test.

Boundaries of declared coverage are also a problem for BigY. Of 6974 SNPs expected to be common to five or more people, 276 (4.0%) are not correctly called. Of these, 51 (0.73%) are “no calls”, 220 (3.2%) are false negatives on the lower end of coverage boundaries, and 5 (0.07%) are false negatives in the main body of coverage. In total, 370 SNP calls are made on lower coverage boundaries, resulting in a 59.5% false negative rate.

A total of 540 SNPs were listed as having incorrect calls in BigY tests. Of these, 177 are “correctly” called SNPs shared by all testers but are listed as inconsistent as they have gaps for no calls and coverage boundaries, leaving 363 SNPs which are sporadically called and ignored (e.g. L128). A total of 3553 false positives are counted, at a rate of one per 1.216 million calls (0.000 082%).

A typical test has 33.13 (st.dev. 6.79) SNPs underneath U106 once all these factors are taken into account, or one per 307,162 base pairs. Including the DYZ19 region would give 36.33 SNPs, or one per 291,393 base pairs, or 5.2% more. An average of 8.71 SNPs per test are estimated to be called sporadically, leading to one SNP per 243,227 base pairs, or 26.3% more SNPs in the raw results as received from Family Tree DNA compared to the final cleaned results.

Of these, 164 SNPs are incoherently called twice, and could represent two instances of a new novel variant. With 33 SNPs per test, lottery mathematics expects one in every 9346 SNPs* to overlap. For 7212 unique SNPs, the probability of this happening once is 54%**. The likelihood is therefore that only one out of these 164 SNPs is actually two instances of a novel variant.

```
     *=COMBIN(101762179;33)/COMBIN(33,1)/COMBIN(101762719-33;33-1)
     **=1-(1-1/[*])^7212
```

CHARACTERISATION OF FGC Y-ELITE

Our team has yet to perform a proper characterisation of the Y-Elite data in preparation for further analysis of it. A space is reserved here for when that is complete.

The history of U106

(1) INTRODUCTION

This deep phylogenic tree of the human population represents our current understanding of the way the human family tree has divided along its male lines. This is a rapidly-evolving field, thus the information is subject to considerable change over time.

This tree summarises the extensive tree that lies above U106. This shows how U106, which now represents many tens of millions of men worldwide, branched off from the rest of the human Y-chromosome tree at different points in prehistory.

(2) OUT OF AFRICA

Ultimately, we all descend from the first life-forms, which lived approximately three billion years ago. Through a long and convoluted process, they evolved into *homo sapiens*. While *H. sapiens* has only been around for about half a million years, this is still older than the common ancestor of the male lines of every person alive today. We call this person Y-chromosomal Adam, because we all descend from him via our father's father's father's father's… etc. Recent estimates of his age vary widely from 120,000 to 380,000 years ago.

The vast majority of people descend through Haplogroup A. In fact, it's only recently that researchers discovered our most-distant relations hiding among remote Africa tribes. Haplogroup BT arose in Africa about 70,000 years ago, when the most of the human population consisted of a small number of tribes living in the Horn of Africa.

The human genetic tree continued to diversify and flourish as mankind expanded throughout Africa. Around 50,000 to 60,000 years ago, a small group of migrants is thought to have crossed the Red Sea into Arabia, starting the most important in a series of Out of Africa migrations.

Some time not too long after this point, a little over 45,000 years ago, we split from haplogroups G and I, which appear to form the original modern human population in Europe. This point is defined by the recently analysed 45,000-year-old remains from western Siberia, from a man who was haplogroup K (but not haplogroup LT).

Our base haplogroup, R, arose from this migration between 24,000 and 34,000 years ago. This is again limited by the archaeological remains of Mal'ta Boy, who was buried 24,000 years ago in Siberia. By this time, our ancestors had probably expanded to across much of north-west Asia, where they existed as hunter gatherers.

(3) EXPANSION INTO EUROPE

Within haplogroup R, most people are part of R1, descended from an individual living 24,000 to 34,000 years ago. The majority of western Europe is descended from the R1 founder. Within R1, there is a bifurcation into two groups: R1a, or M420, and R1b, or M343. R1a is strongest in eastern populations, where it can exceed 60% of individuals in Poland and the south-west Russian states. Its British content is thought to be strongly Viking in origin.

R1b (M343) is thought to have arisen less than 18,500 years ago. In Europe, it is very much dominated by R1b1a2, or M269. This group alone makes up over half the population in Western Europe, and makes up over 90% of some populations. Despite this, its origins are still thought to have been in western Asian populations, and it came to dominate Europe as it expanded throughout the continent.

The date of this expansion into Europe can probably be tied to the sudden growth in the number of branches below M269, which can be very roughly dated to around 4000 BC. The origin of this migration and its route into Europe are not well determined at present. However, archeaological remains show that there was extremely few haplogroup R men in Europe before 2600 BC, when remains from both R1a and R1b are found in Corded Ware and Bell Beaker burials (respectively) in south-eastern Germany.

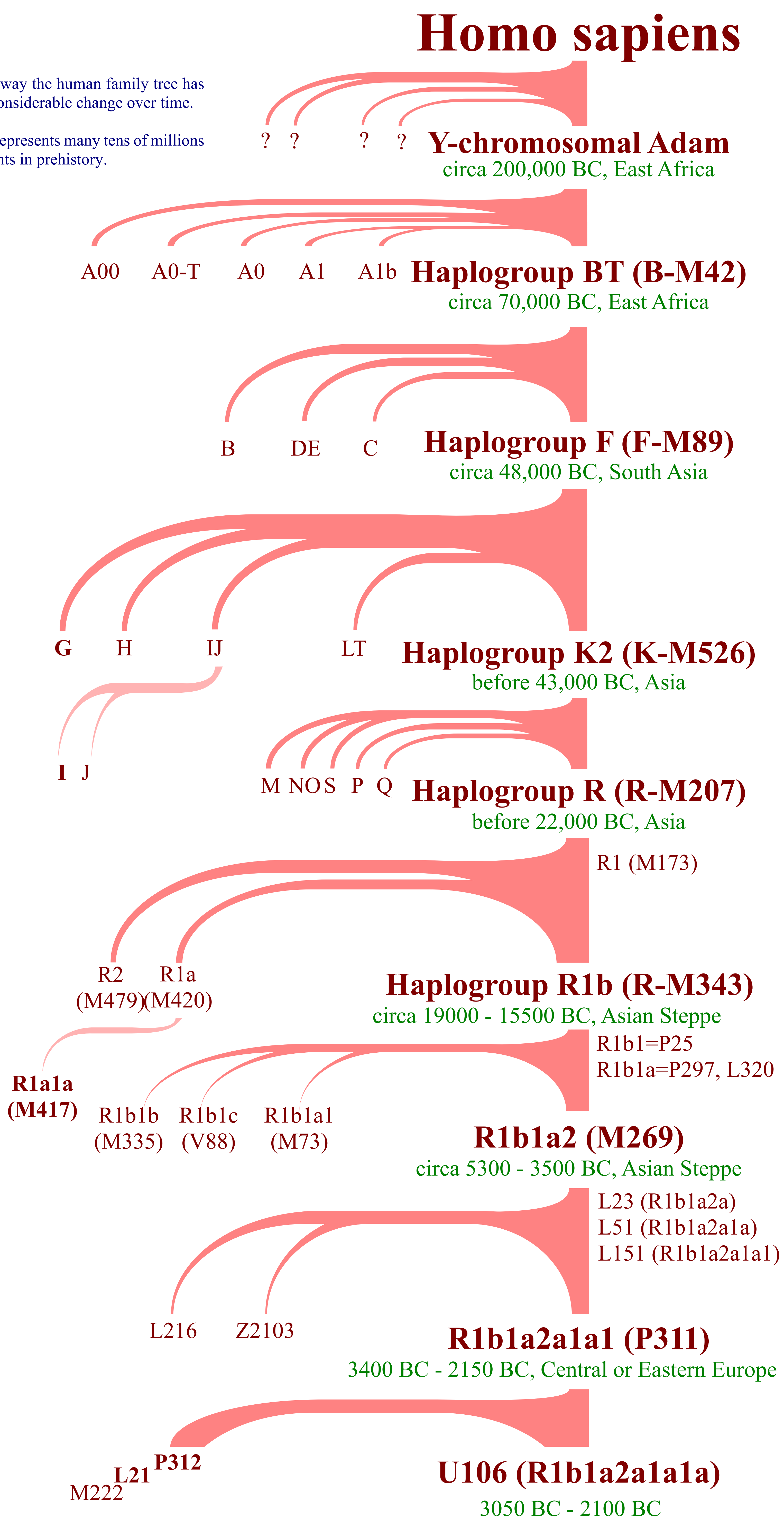
(4) FOUNDING A NEW EUROPEAN POPULATION

Most of the branches above U106 are minor, however there is one important branch at the level immediately above U106, signified by the mutation P311. A split exists at this point in our family tree between the larger P312 branch and the smaller U106 branch.

The P312 branch is generally found more on Europe's Atlantic Coast, while the U106 branch is generally found more in Europe's heartland. This has led to P312 being referred to synonymously with "Celtic" peoples while U106 is "Germanic". While there is clearly some overlap between membership of these SNPs and populations, both SNPs originate several thousand years before these terms are relevant.

Nevertheless, it is the last common ancestor of these two branches, "Mr. P311" whose clan is now represented by around half of western European men, with a third of a billion diaspora worldwide (see panel at right). The date of this man's birth is likely to be during the European Bronze Age, and the possible range of dates correspond to a series of archaeological horizons spreading eastwards over Europe at the same time.

Within P311, U106 represents about 1/8th of Europe, or 110 million men worldwide. We estimate its age to be between 2500 and 4600 years old. We trace what is known about the migrations from Asia to Europe on the next page.



Deep ancestry of U106

Acknowledgements

The information in this tree comes from a variety of sources, but I am most grateful to the International Society of Genetic Genealogy (ISOGG) for maintaining the underlying tree structure displayed here. The anthology of haplogroup statistics on eupedia.com has also been instrumental in creating these data.

Created by: Dr. Iain McDonald; updated: 17 Nov 2014

How to read this chart

This chart shows how the male-line genetic (phylogenic) tree splits from its foundation down to the U106 branch. Different ages and geographical origins distances are shown on the chart, which should be interpreted carefully.

Where quoted, ages are given as 95.5% confidence intervals, what we call "2-sigma". We are 95.5% sure that the real dates lie between these two boundaries. By dividing the uncertaintty in half, we can recover the 68% confidence interval, or "1-sigma" range. Dates are rounded to the nearest 50 years. For example, we are 95.5% sure that the U106 founder lived between 3260 BC and 1974 BC. We are 68% sure that he lived between 2938 and 2295 BC.

This date was calculated using SNP-counting methods which are detailed on later pages.

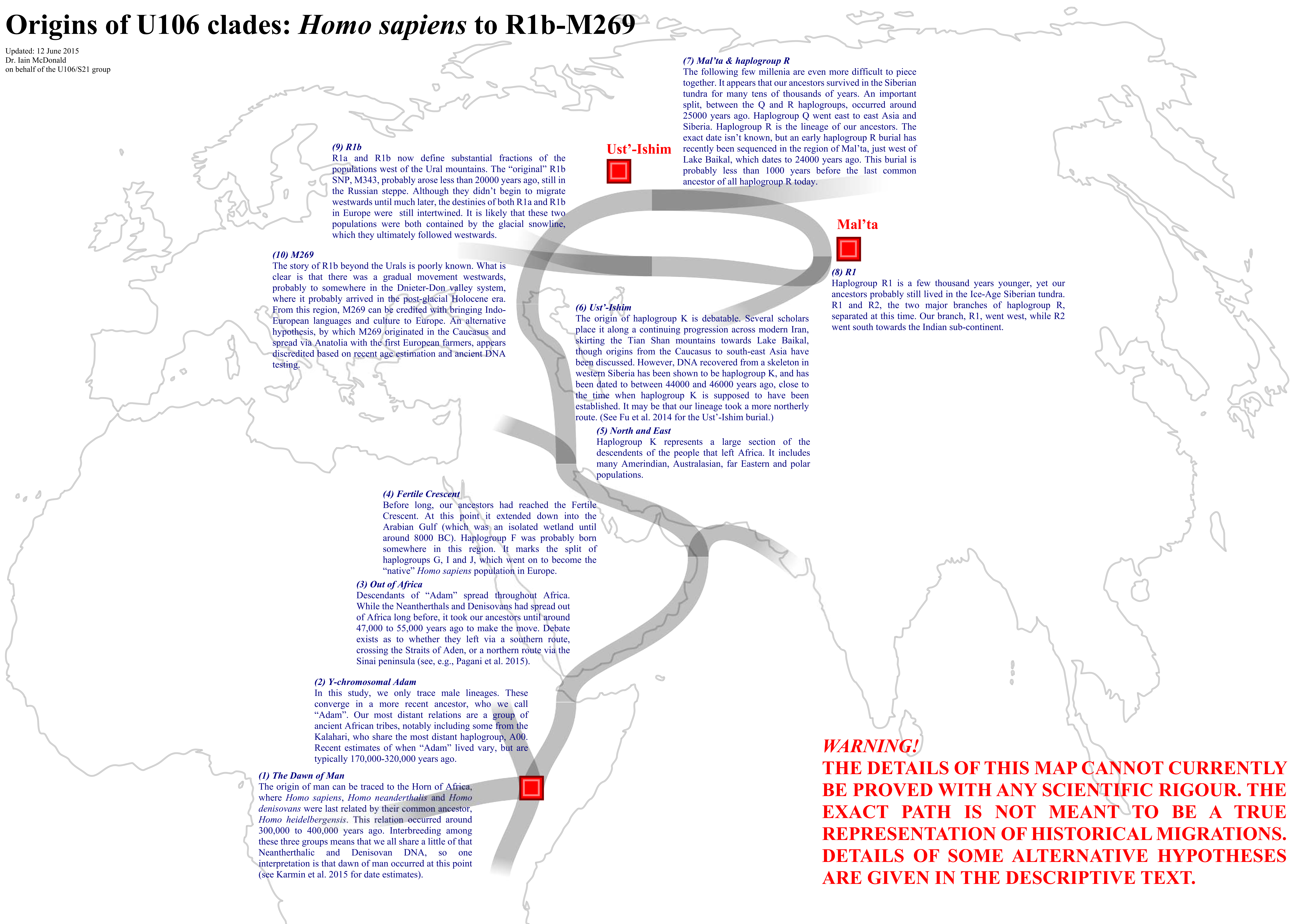
Haplogroup Frequencies in Europe

The following data give the number and percentage of various levels between R1b-M343 and U106 in different parts of Europe, as found by Myers et al. (2007) and selected other studies. These can be used to approximate correction factors to debias our statistics according to how many people of different ancestries have tested. These numbers are only very approximate in many cases and only represent first-order estimates of the underlying population.

COUNTRY	POPLN.	%M269	%U106	M269 & U106 POPLN.		#TESTERS	WEIGHT
British Isles							
Ireland	6429508	80%	6%	5143606	385770	99	4
Scotland	5327000	72.5%	12%	3862075	639240	132	5
England	53012456	57%	20%	30217099	10602491	317	33
Wales	3063456	83.5%	5%	2557985	153172	13	12
Total	*67836420*	*62%*	*19%*	*41780765*	*11780673*	*658*	*18*
Iberia							
Spain	47150800	42%	5%	19803336	2357540	6	629*
Portugal	10607995	56%	5.2%	5940477	551616	3	53*
Total	*57758795*	*45%*	*5%*	*25743813*	*2909156*	*9*	*323**
Scandinavia							
Norway	4930116	25%	15%	1232529	739517	31	24
Sweden	9360113	15%	10%	1404017	936011	29	32
Finland	5357537	2%	1%	107151	53575	8	7*
Total	*19647766*	*14%*	*9%*	*1511168*	*1729103*	*68*	*25*
Central Europe							
Denmark	5568854	34%	17%	1893410	946705	9	105*
Netherlands	16696700	54%	35%	9016218	5843845	32	183
Belgium	11198638	59.5%	25%	6663189	2799659	10	280
France	65460000	52%	7%	34039200	4582200	21	218
Germany	81757600	43%	19%	35155768	15533944	103	151
Switzerland	7785000	58%	13%	4515300	1012050	13	78
Italy	60418711	37%	4%	22354923	2416748	14	173
Austria	8414638	27%	23%	2271952	1935366	2	968*
Total	*257300141*	*45%*	*14%*	*115909960*	*35070517*	*204*	*172*
Eastern Europe							
Hungary	9979000	20%	4%	1995800	399160	6	67*
Czech Rep.	10261320	28%	14%	2873169	1436584	5	287*
Slovakia	5443386	25%	3%	2721693	326603	1	327*
Poland	38192000	23%	8%	8784160	3055360	19	161
Lat./Lit./Est.	6032500	10%	4%	603250	241300	12	20
Belarus	9503807	5%	0.5%	475190	47519	1	48*
Ukraine	45939820	25%	9%	11484955	4134583	4	1034*
Romania	20121641	15%	2%	3018246	402432	1	402*
Bulgaria	7621337	10%	2%	762133	152426	0	-*
Former Yugo.	20449929	5%	1%	1022496	204499	1	204*
Slovenia	2012917	17%	4%	342195	80516	3	27*
Greece	11645343	10%	1%	1164534	116453	0	-*
Russia	110000000	21%	5.4%	23100000	5940000	7	849*
Turkey	76667864	14%	0.4%	10733500	306671	0	-*
Total	*373870864*	*18%*	*4.5%*	*69081321*	*16844106*	*60*	*281*
European Colonies (estimated)							
United States	230000000	46%	15%	105800000	34500000	-	-
Australia	20000000	46%	15%	9200000	3000000	-	-
NZ	4000000	46%	15%	1840000	600000	-	-
Canada	30000000	46%	15%	13800000	4500000	-	-
Total	**1041 million**	**N/A**	**N/A**	**386 million**	**111 million**	***(* Bias factor highly uncertain)***	

Origins of U106 clades: *Homo sapiens* to R1b-M269

Updated: 12 June 2015
Dr. Iain McDonald
on behalf of the U106/S21 group



(1) The Dawn of Man

The origin of man can be traced to the Horn of Africa, where *Homo sapiens*, *Homo neanderthalis* and *Homo denisovans* were last related by their common ancestor, *Homo heidelbergensis*. This relation occurred around 300,000 to 400,000 years ago. Interbreeding among these three groups means that we all share a little of that Neanthertalic and Denisovan DNA, so one interpretation is that dawn of man occurred at this point (see Karmin et al. 2015 for date estimates).

(2) Y-chromosomal Adam

In this study, we only trace male lineages. These converge in a more recent ancestor, who we call "Adam". Our most distant relations are a group of ancient African tribes, notably including some from the Kalahari, who share the most distant haplogroup, A00. Recent estimates of when "Adam" lived vary, but are typically 170,000-320,000 years ago.

(3) Out of Africa

Descendants of "Adam" spread throughout Africa. While the Neanthertals and Denisovans had spread out of Africa long before, it took our ancestors until around 47,000 to 55,000 years ago to make the move. Debate exists as to whether they left via a southern route, crossing the Straits of Aden, or a northern route via the Sinai peninsula (see, e.g., Pagani et al. 2015).

(4) Fertile Crescent

Before long, our ancestors had reached the Fertile Crescent. At this point it extended down into the Arabian Gulf (which was an isolated wetland until around 8000 BC). Haplogroup F was probably born somewhere in this region. It marks the split of haplogroups G, I and J, which went on to become the "native" *Homo sapiens* population in Europe.

(5) North and East

Haplogroup K represents a large section of the descendents of the people that left Africa. It includes many Amerindian, Australasian, far Eastern and polar populations.

(6) Ust'-Ishim

The origin of haplogroup K is debatable. Several scholars place it along a continuing progression across modern Iran, skirting the Tian Shan mountains towards Lake Baikal, though origins from the Caucasus to south-east Asia have been discussed. However, DNA recovered from a skeleton in western Siberia has been shown to be haplogroup K, and has been dated to between 44000 and 46000 years ago, close to the time when haplogroup K is supposed to have been established. It may be that our lineage took a more northerly route. (See Fu et al. 2014 for the Ust'-Ishim burial.)

(7) Mal'ta & haplogroup R

The following few millenia are even more difficult to piece together. It appears that our ancestors survived in the Siberian tundra for many tens of thousands of years. An important split, between the Q and R haplogroups, occurred around 25000 years ago. Haplogroup Q went east to east Asia and Siberia. Haplogroup R is the lineage of our ancestors. The exact date isn't known, but an early haplogroup R burial has recently been sequenced in the region of Mal'ta, just west of Lake Baikal, which dates to 24000 years ago. This burial is probably less than 1000 years before the last common ancestor of all haplogroup R today.

(8) R1

Haplogroup R1 is a few thousand years younger, yet our ancestors probably still lived in the Ice-Age Siberian tundra. R1 and R2, the two major branches of haplogroup R, separated at this time. Our branch, R1, went west, while R2 went south towards the Indian sub-continent.

(9) R1b

R1a and R1b now define substantial fractions of the populations west of the Ural mountains. The "original" R1b SNP, M343, probably arose less than 20000 years ago, still in the Russian steppe. Although they didn't begin to migrate westwards until much later, the destinies of both R1a and R1b in Europe were still intertwined. It is likely that these two populations were both contained by the glacial snowline, which they ultimately followed westwards.

(10) M269

The story of R1b beyond the Urals is poorly known. What is clear is that there was a gradual movement westwards, probably to somewhere in the Dnieter-Don valley system, where it probably arrived in the post-glacial Holocene era. From this region, M269 can be credited with bringing Indo-European languages and culture to Europe. An alternative hypothesis, by which M269 originated in the Caucasus and spread via Anatolia with the first European farmers, appears discredited based on recent age estimation and ancient DNA testing.

WARNING!
THE DETAILS OF THIS MAP CANNOT CURRENTLY BE PROVED WITH ANY SCIENTIFIC RIGOUR. THE EXACT PATH IS NOT MEANT TO BE A TRUE REPRESENTATION OF HISTORICAL MIGRATIONS. DETAILS OF SOME ALTERNATIVE HYPOTHESES ARE GIVEN IN THE DESCRIPTIVE TEXT.

Age estimation

AGE ESTIMATION FROM BIG Y

The formation of SNPs is a largely random process. Many processes affect genetic integrity and structure, many carcinogens cause genetic mutations (harmful or otherwise), and many social and environmental factors affect the number of mutations passed from father to son. However, these largely cancel out when one considers a large population over a long time. Certainly, SNP creation seems like a random process within the errors of our observations.

SNP mutations can therefore act as a clock, albeit one that does not have a regular tick. SNP creation is a roll of the dice: sometimes you will get one, sometimes you won't. Sometimes it will be in your tested region, sometimes it won't. Over long timescales, and many lineages, these effects cancel out, so that there is a particular rate at which SNPs form. We can therefore expect that SNPs will build up in tests like BigY at the rate of a certain number of years per SNP, which we will call *r*.

At its simplest, the age of a clade (*t*) can be estimated by the taking the number of SNP mutations that are not shared by all the members of that clade (*m*), multiplying it by the timescale for SNP formation (*r*) and dividing it by the number of testers (*n*), thus:

$$t = m\, r \,/\, n$$

where *m* and *n* come from the BigY tests, and *r* comes from some nominal, independent measurement. For large clades, the rate *r* is the most uncertain parameter in this calculation: *n* is known precisely, and *m* is typically determined to much better than 3%. For small clades, the fact that SNP creation is a random process becomes important, and the small-number statistics of *m* are the dominant uncertainty.

ACCOUNTING FOR SMALL-NUMBER STATISTICS

Small-number statistics of SNP creation is governed by a branch of mathematics called Poisson statistics. Poisson statistics tells us the probability of observing any given number of mutations in a single lineage, compared to what a regular mutation rate "clock" would give. We reverse-engineer this calculation to find the uncertainty in the number of mutations we see (δm).

For large clades, calculating this uncertainty becomes technically impractical, so we use the Gaussian approximation that the 1-σ (68.3%) uncertainty in *m* is the square root of *m*, and that the 1.96-σ (95%) uncertainty is 1.96 × $\sqrt{m}$.

An additional uncertainty comes from the conversion of SNPs to years. This is because the mutation rate comes with its own uncertainty (δr). Since these uncertainties are uncorrelated, they are added in quadrature, such that:

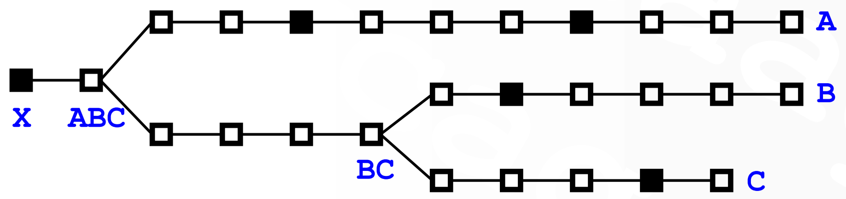
$$\delta t/t = \sqrt{([\delta m/m]^2+[\delta r/r]^2)}$$

This gives the age and its uncertainty listed in the final age products shown in this work, and is the final way in which the age of U106 is worked out.

TMRCA vs. BRANCH AGE vs. SNP AGE

What this calculation gives you is the time between the birth of the most-recent common ancestor and the average birth date of the *n* testers which have been tested. This is the "time to most-recent common ancestor" or **TMRCA**. This is subtly different from the **SNP age**: the actual age of the quoted SNP. In most cases, this distinction doesn't matter, but it can become important in some clades.

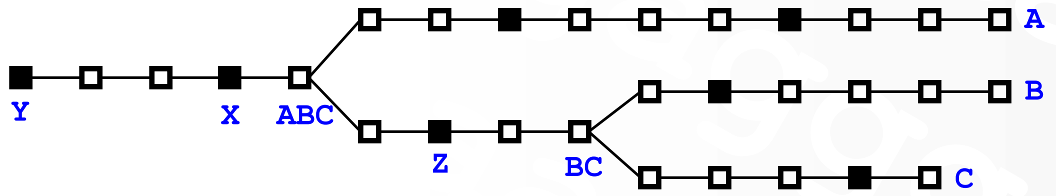
In the simplest case, we might have the following family tree, where every box represents the birth of a son and filled boxes represent the creation of a new SNP:



In this case, A, B and C share a most-recent common ancestor (ABC) and a terminal common SNP (X). The age of X is slightly older than that their TMRCA, but this can usually be ignored.

However, if only B & C take a next-generation test, and their common ancestor (BC) has not had any further mutations since the ABC ancestor, the TMRCA for B & C might be a century or two younger than either ABC or X.

This can become more serious if we have the following scenario:



Here, B & C share a set of SNPs (X,Y and Z). If only B & C test, we would get the following test results:

```
B: X+ Y+ Z+, 1 novel variant
C: X+ Y+ Z+, 1 novel variant
```

we have no idea which one out of X, Y or Z comes first. These lists of SNPs can become very long (30 or more SNPs), so they are often abbreviated by one of the SNPs, in this case "X". So what we write is the age of X (because we do not know any better), but what we calculate is the time since the birth of BC.

If tester A then comes along with the results:

```
A: X+ Y+ Z-, 2 novel variants
```

then we will know that X and Y come before Z. The recorded age of X will change, as the common ancestor ABC is much older than BC. It is therefore important to bear in mind the fact that what we are reporting is the time since the birth of the most-recent common ancestor of all people with the indicated SNP ***who have taken a BigY test***.

Sometimes additional data is available (e.g. from Y-Prime, Y-Elite, or individual SNP testing at YSeq or Family Tree DNA) that can split long chains of SNPs in this fashion. This data is not included when calculating the ages above, as it is not homogeneously reduced.

A MORE ACCURATE AGE

Particularly in the case of small clade branching off from a much larger one (e.g. S5520 under Z156 or FGC396 under U106), a more accurate age can be derived by considering the time between the parent SNP and the target SNP.

This can be done in a similar manner, considering the number of SNPs between the parent and target SNP (m_p). This provides a more accurate answer when m/n is much larger than m_p. Excluding the DYZ19 region, for FGC396's two testers Lindemann and Kuykendall, $m_p = 7$ while $m/n = 17$. In practice, we can do this both from the U106 age and from the age of the immediate parent SNP, as sometimes one is more accurate than the other.

A final modification we can make is based on this method. If we fix the age of U106 using our original method, then we can adapt the ages for the fact that some lines (e.g. L48 averages 36.41) have more mutations than average, while some (e.g. Z18 averages 27.04) have fewer. This difference is expected, as larger clades will preferentially have more SNPs due to random sampling. This is exemplified in the two trees presented earlier, where the first tree produces three small clades, but the addition of SNP "Z" produces two clades, of which clade Z is larger. This is particularly effective during population expansion periods.

In this final method, we have a fixed age of U106 (let's say it's 4500 years). If we have a clade under U106 with an average of 45 SNPs, we can fix a mutation rate for this lineage of one SNP per 100 years. If it has an average of 22.5 SNPs, it will be one per 200 years. Naturally, our uncertainty measurement has to take this new mutation rate and its uncertainties into account.

Using these methods, we have a suspension-bridge-like design, whereby the origin of the tree, U106, is fixed from the present day. Clades are pinned to this tree both downwards from U106 via their parent lines, and up from the present day. The intersection of these two methods provides much more stable and self-consistent ages for each SNP than would be arrived at otherwise.

AGES OF INDIVIDUAL SNPS

Ages of the actual SNPs are more uncertain, given the processes described above. However, they will occur at a fixed time before the TMRCA or convergence age. This is given by:

$$t_s - t = r\,(n_r - 0.5)\,/\,2$$

where n_r is the number of SNPs in an unbroken run (e.g. Z305, Z306, Z307, S1667 would give $n_r = 4$).

The 95% uncertainty on this is again computed from Poisson statistics, but asymptotes to +/- 0.475 $r\, n_r$ for large n_r.

FINAL AGE CALCULATION

The final age is determined from three numbers:
Firstly, from the number of SNPs beneath the target:

$$t = m\ r\ /\ n \qquad \textit{[T1]}$$

Secondly, from the number of SNPs between U106 and the target:

$$t_0 = t(\mathrm{U106}) - m_0\ r_0 \qquad \textit{[T2]}$$

where t(U106) is the age of U106 from ***[T1]*** and m_0 is the number of mutations since U106. Here, r_0 is defined from the average number of mutations in that branch since U106 (m(U106)) as follows:

$$r_0 = m(\mathrm{U106})\ /\ t(\mathrm{U106}) \qquad \textit{[R2]}$$

Thirdly, from the number of SNPs since the parent clade:

$$t_\mathrm{p} = t_0(p) - m_\mathrm{p}\ r_0 \qquad \textit{[T3]}$$

where $t_0(p)$ is the age of the parent from ***[T2]*** and m_p is the number of mutations between the parent and the target SNP. ***[T1]*** can be adapted for a given clade such that:

$$t = m\ r_0\ /\ n \qquad \textit{[T4]}$$

which then gives the equality:

$$t = t_0 = t_\mathrm{p} \qquad \textit{[T5]}$$

such that ages from the three estimates are consistent. A final modification to this age is made in the rare case that a sub-clade has a larger average number of SNPs beneath it than its parent ($m/n > m_\mathrm{p}/n_\mathrm{p}$). In this case, a hard limit are placed of at least 30 years after the parent clade's origin. A hard limit is also placed at 1950, representing an age of zero.

FINAL AGE UNCERTAINTY

The uncertainty in the final age estimation is a combination of the uncertainties derived from equations ***[T2]***, ***[T3]*** and ***[T4]***. It therefore relies on the uncertainty in the U106 age. For a 95% confidence interval, this is the 1.96-σ uncertainty value, namely:

$$\delta t/t = 1.96\ \sqrt{([\delta m/m]^2+[\delta r/r]^2)} \qquad \textit{[ET1]}$$

where δr is derived from the literature or case studies. Similiarly, the uncertainty in ***[T4]*** can be derived as:

$$\delta t/t = 1.96\ \sqrt{([\delta m/m]^2+[\delta r_0/r_0]^2)} \qquad \textit{[ET2]}$$

where δr_0 is given from ***[R2]*** by:

$$\delta r_0 = 1.96\ [m(\mathrm{U106}) -/+ \delta m(\mathrm{U106})]\ /\ [t(\mathrm{U106}) +/- \delta t(\mathrm{U106})] \quad \textit{[ER1]}$$

In both cases, $m-\delta m$ and $m+\delta m$ are given by the highest and lowest value of λ, respectively, for which:

$$\int_{k=0}^{k=m} \mathrm{Pois}(\lambda,k)<0.1585 \qquad \textit{[ER2]}$$

$$\int_{k=0}^{k=m} \mathrm{Pois}(\lambda,k)>0.8415 \qquad \textit{[ER3]}$$

at 1σ and:

$$\int_{k=0}^{k=m} \mathrm{Pois}(\lambda,k)<0.025 \qquad \textit{[ER4]}$$

$$\int_{k=0}^{k=m} \mathrm{Pois}(\lambda,k)>0.975 \qquad \textit{[ER5]}$$

at 95% confidence, where Pois() is the Poisson function, $(\lambda^k/k!)e^{-\lambda}$. For large m, where this value is computationally expensive to determine, the approximation $\delta m = 1.96\ \sqrt{m}$ is used for the 95% confidence interval.

The uncertainty in the other two age measurements follows similar principles, except that the uncertainty in m_0 and m_p replaces the uncertainty in m, and the age is calculated in time since ($t\pm\delta t$) for U106 and the parent SNP, respectively, rather than as an age from the present day.

If $\delta t_0 < \delta t_\mathrm{p}$ (i.e. the age from U106 is more accurately determined than the age from the parent clade), then the U106-based age is used, otherwise the age based on the parent SNP is used. This provides an age propagated forward in time, which we will call t_e with associated uncertainty δt_e. Note that because ***[ER4]*** and ***[ER5]*** do not provide errors symmetric around m, the final uncertainty, δt_e, will not be symmetric around t either.

The age uncertainties can be combined using a weighted average to produce a final uncertainty in the convergence age as follows:

$$\delta t_\mathrm{final} = \frac{\left(\dfrac{t\pm\delta t}{w} + \dfrac{t_\mathrm{e}\pm\delta t_\mathrm{e}}{w_\mathrm{e}}\right)}{\left(\dfrac{1}{w} + \dfrac{1.}{w_\mathrm{e}}\right)} \qquad \textit{[ET3]}$$

where the weights are set as follows:

$$w = (t/n\ .\ 2\delta t)^2 \qquad \textit{[ET4]}$$

$$w_\mathrm{e} = ([t^* - t_\mathrm{e}\ ].\ 2\delta t_\mathrm{e})^2 \qquad \textit{[ET5]}$$

where t^* is either t_p or t(U106), depending on which provides the more accurate age. The same limits are applied such that the cluster cannot be older than its parent and cannot be younger than the present day.

DEFINING THE PRESENT DAY

In this work, we use 1950 as being the present day, representing the average birth date in the testing population. This comes from an online survey of 98 DNA testers from the U106 group itself. The average birth year of these testers is 1950.3 with a standard deviation of 15.5 (i.e. a 1.96-σ uncertainty of 30.4 years for a single tester or 1.50 years for the total BigY testing population).

This estimate is likely to be slightly biased by those individuals who are active on the online forum compared to the underlying dataset, but overall this is expected to impart a relatively small uncertainty to the age of any particular SNP.

CHOOSING A MUTATION RATE

We have so far ignored how the choice of the underlying mutation rate, r, and its uncertainty, δr, are calculated. Ultimately, these come from literature studies which sum up a measured number of mutations which have occurred over a known period of time.

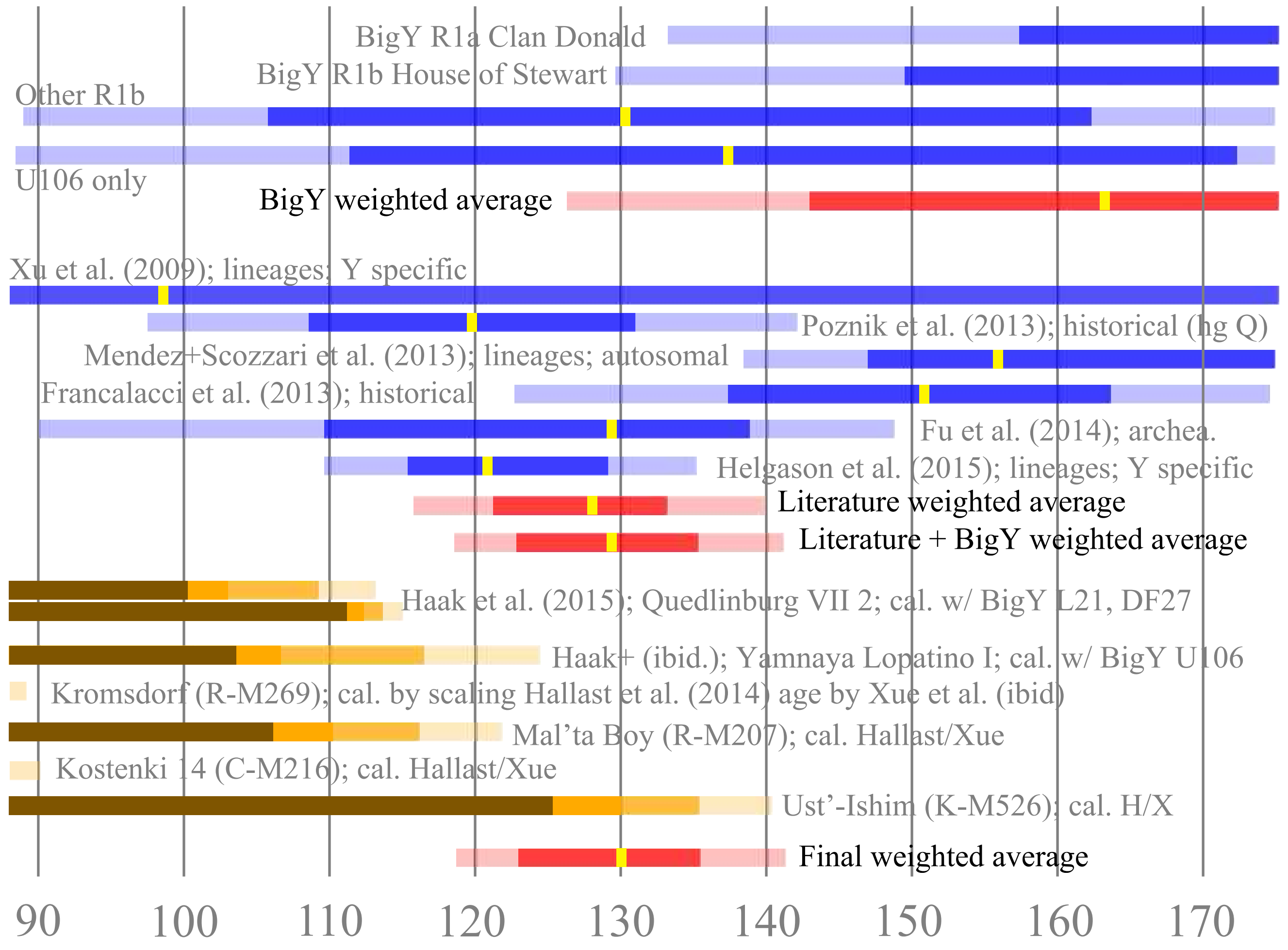
At their best, these are studies of large lists of known genealogies, where the years between each father and son are added up, along with the mutations that have accumulated during that time. The ratio of these directly gives the mutation rate. Some studies give this as an average ratio across all chromosomes, some assume the Y-chromosome rate is the same as the autosomal rate. Most studies treat the Y-chromosome as a single, uniform entity, although some have split it up into regions with different characteristics.

In lieu of this detailed study, isolated populations with well-determined archaeological convergence dates are often modelled. These rates are not direct measurements and are subject to genetic drift, archaeological dating uncertainties and models of population size evolution. However, they provide a constraint on any differences between ancient DNA mutation rates and modern ones, and typically cover well the timescales of interest.

A final constraint comes from archaeological DNA. In general, these only provide a lower limit to the mutation rate. If a sample is of a known haplogroup (e.g.) and is of (at least) a given age, the number of mutations since that haplogroup was formed over the age of the sample provides a limit to possible fast mutation rates. If an estimate can be made of the number of mutations that sample has had since the haplogroup formed, this can also constrain slow mutation rates.

In the next section, are shown a list of rates found in the literature, as applied to BigY. Full notes on their methods and homogenisation are detailed in the supplementary information in the associated file (snp-mutation-rate.xls) on deposit in the U106 forum or available on request.

MUTATION RATES



The above chart shows the mutation rate in years per SNP per BigY test, as limited by known lineages in the BigY itself, literature rates and archaeological data. It can be read as follows:

- Blue lines show the results from individual studies. The yellow point marks the best-estimate value, and the shaded blue regions show the 68.3% and 95% confidence intervals.
- Red lines show the weighted average of several results. Here, the square of the confidence range is used as a weight.
- Orange lines show the limits obtained from archaeological data. The different shadings (darker→lighter) show the regions ruled out at 99.75%, 95%, 68.3% and 50% confidence.

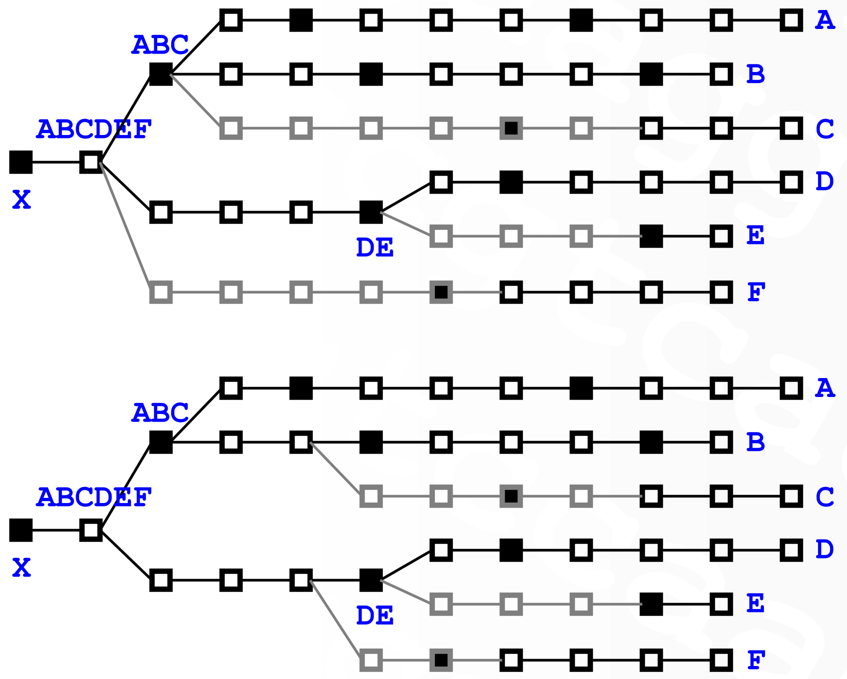
In the following, we explore the rates shown above.

RATES FROM LINEAGES IN BIGY

As of May 2015, we have 91 BigY tests from lineages where we have a named individual who is very likely the common ancestor of at least two tests. "Very likely" in this case is a judgement call made based on paper-trail and genetic evidence. In total, they represent around 40,000 years of lineages and produce an average rate of 163 years/SNP (95% c.i.: 126-208 years/SNP).

These are dominated by two Scottish families: the Clan Donald and the House of Stewart. In most cases with these lineages, we do not have the complete paper trail leading from the testing individual back to the common ancestor, but we have other BigY tests which show that they must come from this lineage.

These cases suffer from problems in accounting for the number of years in a lineage. In the following figure, we consider six testers (A through F) of which we have full paper trails from A, B and D. Paper trails from C, E and F are only partially known (shown in gray). We show two possible configurations for the family tree, depending on whether C, E and F branch earlier or later. As before, black squares denote generations in which SNPs occur. This figure is a simplification of the situation in the House of Stewart.



In either case, the total number of years can be found by summing the lengths ABCDEF→ABC + ABC→A + ABC→B + ABC→C + ABCDEF→D + ABCDEF→E + ABCDEF→F. However, the uncertainties are larger than for a family where we know the entire family tree. We can better account for structure we know (e.g. the relationship between D & E is fixed by the SNP at DE) than for structure we can't (e.g. the relationship between D & F). This can lead to a systematic bias towards a higher number of years/SNP for large families if there is a long period between ABCDEF and DE where no SNPs occur. It is suspected that this is the reason that the Clan Donald and House of Stewart results give comparatively large rates for BigY tests. Note that these extra uncertainties are not fully accounted for in the previous figure.

BigY tests from other families only account for around 10,000 years of lineages. Although the uncertainties on these are larger, they roughly show the same ~130 years/SNP as the bulk literature.

RATES FROM THE LITERATURE

The only studies the perform a thorough, Y-specific mutation rate estimate are those of Xu et al. and Helgason et al. Helgason et al. additionally provides two estimates, for the palindromic and non-palindromic regions, respectively, which correspond to 113 and 133 years/SNP in BigY. The slower palindromic rate is consistent with the paternally transmitted autosomal rate. This may help why the rates from Mendez et al. (2013) and Scozzari et al. (2013) are higher, as they are scaled from autosomal values.

The archaeological DNA results in the figure have had an extra 10-20% uncertainty added to them to reflect the uncertainty in the date at which the tested population formed. They produce very different values (120/151 years/SNP) which partly reflect this uncertainty.

The Fu et al. (2014) result is based on sequencing from the Ust'-Ishim burial in western Siberia. The modelled age is consistent with the Helgason et al. value and indicates that the mutation rate has not changed significantly over tens of millenia.

A weighted average of the mutation rates from BigY and the literature produce a well-constrained value around 129 years/SNP, with an uncertainty of around 9% that is transmitted directly into the ages of each SNP.

RATES FROM ARCHAEOLOGICAL REMAINS

Obtaining rates from archaeological remains depends on having a known date for the archaeological remains, a known haplogroup for those remains (and preferably a good idea of how long it was between the formation of the haplogroup and the individual's lifetime), and the average number of SNPs formed since that haplogroup's formation in present-day lineages.

The previous figure shows two burials analysed by Haak et al. (2015) which are sub-clades of R-M269. We can use the results from BigY to directly determine the number of SNPs since R-M269 (knowing the upstream number of SNPs between M269 and U106, etc.). This provides a limit (with 50% confidence) that the mutation rate is slower than 124 years/SNP (113 years/SNP at 95% confidence).

Four more ancient burials are also shown. Here, we do not have a clear idea of the number of SNPs that show up in BigY, so we have scaled the mutation rate from Xue et al. (2009) by the ratio of the ages from Hallast et al. (2014) and radio-carbon dating of the remains (Hallast et al. use the Xue et al. rate). While these initially appear more limiting, the Hallast et al. study does not calculate its dates directly from the number of SNPs, but by the rho statistic, hence these limiting rates are indicative only, and should not be rigorously applied. Note that the Fu et al. study mentioned above uses the Ust'-Ishim burial, and arrives at a much faster mutation rate than is apparently allowed.

FINAL MUTATION RATE

The mutation rate that is finally used in this document and elsewhere in the U106 group's output is a weighted combination of the BigY and literature results, limited by the archaeological remains.

To create this limit, we assume that the probability distributions from BigY and the literature are Gaussian on either side of the mean (though not necessarily the same Gaussian on either side). We convolve this with the probability distribution from the archaeological literature at the 50%, 68.3% and 95% confidence intervals. This results in a rate which (as of 29 May 2015) is 130 years/SNP, with a 95% confidence interval of 118‒149 years/SNP. This corresponds to 7.56 (6.95‒8.32) x 10^{-10} SNPs per base pair per year.

CALIBRATING STR TO SNP MUTATION RATES

STR markers also seem to behave like a randomly ticking “clock”, so in principle these can be used for age measurements as well. The advantage of using STR markers is that, typically, more people within a clade will have tested for these.

STR dates also provide us with some difficulty. They mutate up and down at a much faster rate than SNPs, so 111 STR markers provides about the same mutation rate as the SNPs in a 10-million-base-pair BigY test. This means that mutations back to the ancestral state are a problem. They can also mutate by more than one step at a time, and the decision has to be made as to whether to count this as one mutation or several. Finally, they also seem to prefer specific values, so will preferentially mutate to these lengths.

This makes the concept of STR dating much more mathematically complicated than SNP dating. Over timescales of a few hundred years, the above problems are negigible, but on longer timescales they become very significant. Usually an exponential multiplier is used to correct STR dates to SNP dates. In the following, we tie the STR age to the SNP ages within U106 using such a scaling relation.

To begin, we discuss methods of age calculation. Each relies on setting an mutation rate for each STR, the source of which we will discuss later.

Infinite allele model: The infinite allele model assumes any variance in an STR is a single mutation, e.g. it treats 15→17 as one “multi-step” mutation, whereas it is possible it was really two mutations: 15→16→17. Generally speaking, the infinite allele model will be more accurate for young clades. For old clades where more than one mutation is likely on some STRs, the step-wise allele model is better.

Step-wise allele model: The step-wise allele model assumes each repeat of an STR is a unique mutation. It counts mutations like 15→17 as two mutations: 15→16→17. Often a hybrid is used which assumes step-wise for all STRs except the multi-copy markers, which are infinite. We do not consider the step-wise model further here.

Variance-based model: Both the step-wise and infinite allele models do not correctly account for back mutations, e.g. 15→16→15. The variance-based method accounts for them in part by taking the mathematical variance of a group of DNA tests, rather than simply counting the mutations.

The infinite allele model we use derives from Dean McGee’s tool (http://www.mymcgee.com/tools/yutility.html), which calculates the time to most recent common ancestor (TMRCA) for a grid of individuals. This data can then be combined using the method described below.

The variance-based model is based on a method and tool developed by Ken Nordtvelt, which has undergone substantial modification to include error estimates and include a number of easily changeable options.

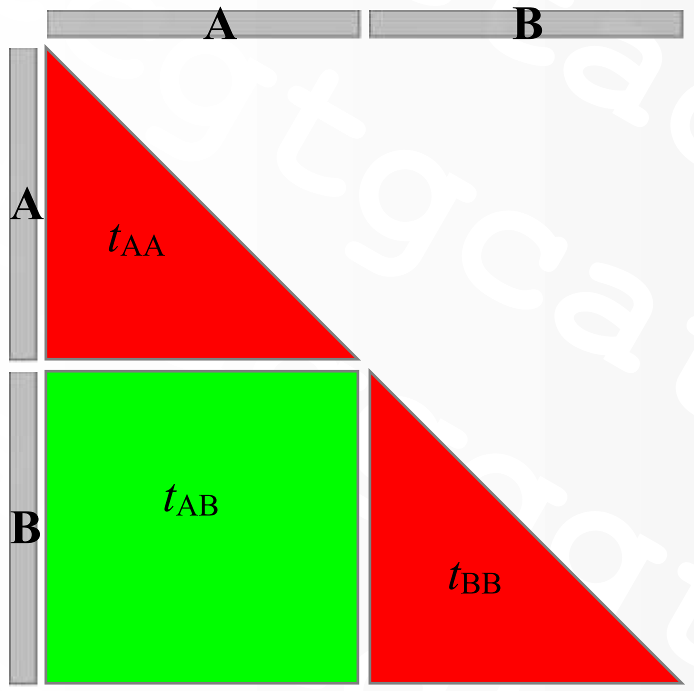
INTRA-CLADE AND INTER-CLADE AGES

McGee’s tool calculates TMRCA for two individuals. What we require is the TMRCA for an entire group. In combining age estimates, it is important to consider whether you want the age within a group (the *intra*-clade age) or the age between two groups (the *inter*-clade age): e.g., do you wish to know the relationship between people who are U106+, or the age when Z18 and Z381 last shared a common ancestor?

Intra-clade ages are generally problematic, as they ignore the fact that many people within a clade are closely related: e.g., many calculations of the intra-clade age of U106 will be biased by the fact that half of people are L48+. The calculated age will be pulled down towards the L48 age. For this reason, inter-clade ages are generally used for STR calculations, which compare two clades to each other.

INFINITE AGE COMBINATIONS

For this method, the McGee tool outputs a tabulated matrix of TMRCAs. Assuming that clade “A” is listed at the top and clade “B” is listed at the bottom, the intra-clade TMRCAs of A and B (t_{AA},t_{BB}), and the inter-clade TMRCA of A and B (t_{AB}) will be given from the intersection of these two sets, which will fall in this region of the table:



Either the average or median value can be taken here as an estimation of the TMRCA of the A–B relationship, and the sample standard deviation can be taken as the standard error on this value. On top of this, there will be a systematic error to account for the uncertainties in the mutation rates, and the dataset must be calibrated against the SNP rates to account for non-random elements in the mutations.

The final age is therefore given by:

$$t_{AB} = \frac{t_{A1-B1} + t_{A1-B2} + \ldots + t_{A1-Bn} + t_{A2-B1} + \ldots + t_{Am-Bn}}{mn}$$

where there are m tests from clade A (A1 through Am) and n tests from clade B (B1 through Bn). The uncertainty is given by:

$$\delta t^2_{AB} = \frac{\sigma(t_{AB})^2}{mn} + \frac{\sigma(\Sigma_{i=1}^{111}\mu_i w_i)^2}{(\Sigma_{i=1}^{111}\mu_i w_i)^2}$$

where $\sigma(t_{AB})$ is the standard deviation among all TMRCAs in the A–B set, μ_i is the mutation rate on marker i and w_i represents a weighting factor which is the fraction of test pairs which are compared on marker i. The left-hand ratio therefore represents the square of the standard error in the mean, and the right-hand ratio represents the square of the fractional uncertainty in the mutation rate. The square root of this gives δt_{AB}, the uncertainty in t_{AB}.

VARIANCE-BASED AGE CALCULATION

Each marker i in test j returns an allele value $x_{i,j}$. The variance among m and n tests in clades A and B, respectively, can be calculated as:

$$\mathrm{Var(AB)}_i = s_{2,A}/m + s_{2,B}/m - 2\, s_{1,A}\, s_{1,B} / mn\ ,$$

where $s_{1,A} = \Sigma_{j=1}^{m} x_{i,j}$, and $s_{2,A} = \Sigma_{j=1}^{m} x^2_{i,j}$, and similarly for $s_{1,B}$ and $s_{2,B}$ for $j = 1$ to n. The square of the fractional uncertainty in that variance (at the 68% confidence interval) will be:

$$\sigma^2(\mathrm{Var(AB)})_i / \mathrm{Var(AB)}^2_i = 2\,(m-1)m^{-2} + 2\,(n-1)n^{-2}\ .$$

Variances on individual markers can be summed, such that:

$$\mathrm{Var(AB)} = \Sigma_{i=1}^{111}\, \mathrm{Var(AB)}_i$$

with a 68% c.i. fractional uncertainty of:

$$\sigma(\mathrm{Var(AB)})/\mathrm{Var(AB)} = \sqrt{}[\Sigma_{i=1}^{111}\, \sigma^2(\mathrm{Var(AB)})_i / \mathrm{Var(AB)}^2_i] / 111$$

Using a mutation rate for each marker, μ_i, the age of the clade can be deduced by:

$$t(\mathrm{AB}) = \mathrm{Var(AB)}\ \Sigma_{i=1}^{111}\mu_i / 2$$

and:

$$\sigma(t(\mathrm{AB})) = t(\mathrm{AB})\ \sqrt{}\{[\sigma(\mathrm{Var(AB)})/\mathrm{Var(AB)}]^2 + [\sqrt{}\Sigma_{i=1}^{111}\,\sigma^2(\mu_i)]^2\}$$

where $\sigma(\mu_i)$ is the (68%) uncertainty in μ_i.

YEARS PER GENERATION

Conventionally, STR mutation rates are given in mutations per generation, whereas we need mutations per year. The conversion of years per generation has adopted many values between 20 and 40 years/gen in the literature. The value varies over time and over societies. Historical studies of populations (particularly in Iceland) indicate it is likely to have been around 35 years/generation over the 16th to 19th Centuries. Since then, a series of scientific and social revolutions have decreased the years/generation (19th Century sanitation improvements, 20th Century medical improvements, birth control) and subsequently increased it again (women’s lib. and two-career families).

For pre-modern agrarian communities between 1000 AD and the present, we adopt 35 +/– 3 years/generation (at 95% confidence). For earlier times, we adopt a scaling that drops to 33 +/– 3 years/gen for 1-1000 AD, 32 +/– 3 years/gen for 1000-1 BC, and 31.5 +/– 3 years/SNP before 1000 BC. Throughout, we adopt a zero point of 1950 AD, +/- 15.5 years at 95% confidence.

CHOICE OF MARKERS

There are various reasons why certain markers may be avoided. These include multi-copy markers like DYS464, where we cannot always tell which value belongs to each copy (e.g. 15-16-17-18 could be a=15, b=16, c=17, d=18 or a=18, b=16, c=17, d=15). We might also select only slowly-mutating markers to select against non-random elements in the mutation process. One final possibility is to use q values (a measure of closeness to random mutation; Bird et al. 2012) to select only STRs that mutate in a close-to-random fashion.

CHOICE OF MUTATION RATES

A variety of mutation rates exist in the literature, with a substantial range in mutation rates. We consider a number of rates here. In the table, the mutation rate source is listed, along with the number of markers contained, the number of those markers used in the following analysis, and the relative mutation rate compared to the average of the ensemble for the markers sampled, where larger numbers indicate faster mutations.

Chandler (2006)	67	50	78%
Doug McDonald (unpub.)	80	not used	247%
Charles Kerchner (unpub.)	67	not used	304%
SMGF	30	not used	109%
FTDNA	37	not used	268%
SMGF/Y-Search	21	not used	117%
Y-HRD	16	not used	83%
Vermeulen et al. (2009)	8	not used	143%
Marko Heinila (unpub.)	111	94	78%
Ballantyne et al. (2010)	91	82	118%
Burgarella et al. (2011)	84	83	118%

We have chosen the indicated four datasets on the basis of number of markers covered and consistency with the average STR mutation rate.

The standard deviation of these four rates over the square root of the average number of rates per marker (3.18) approximates the uncertainty in the rate itself, which we take as our standard (68% c.i.) uncertainty. Overall, this yields a 8.8% systematic uncertainty in the total mutation rate and typically a 5.6% statistical uncertainty in the resulting age at a 68% confidence interval. For comparison, all sources of systematic and statistical error typically yield at least a 11% (21%) uncertainty in the age in generations or a 14% (28%) uncertainty in the age in years at the 68% (95%) confidence intervals.

CONSTRAINTS APPLIED IN THE FOLLOWING

The constraints listed below were applied to the analysis that follows.

- No multi-copy markers were used in any calculation. This leaves 94/111 markers.
- No selection with mutation rate was made unless noted. Where noted, markers with μ>0.004 per generation were discounted unless stated otherwise (leaving 78 markers).
- No selection with Bird's q were made unless noted. Where noted, markers with Bird's q>0.05 were discounted unless stated otherwise (leaving 40 markers).

The 94 markers used give a mutation rate of μ=0.315+/-0.028 per generation, or once per 111+/-14 years.

DATA USED IN THE ANALYSIS

The U106 group's STR database was sampled on 3rd June 2015, and includes 2058 STR entries. Of these, 1722 have a relatively secure placement in a clade under U106, with 141728 markers in total, or an average of 82.3 markers each.

CALIBRATION OF STR TO SNP AGES: VARIANCES

STR ages are usually boot-strapped to SNP ages using some variation on the following expression:

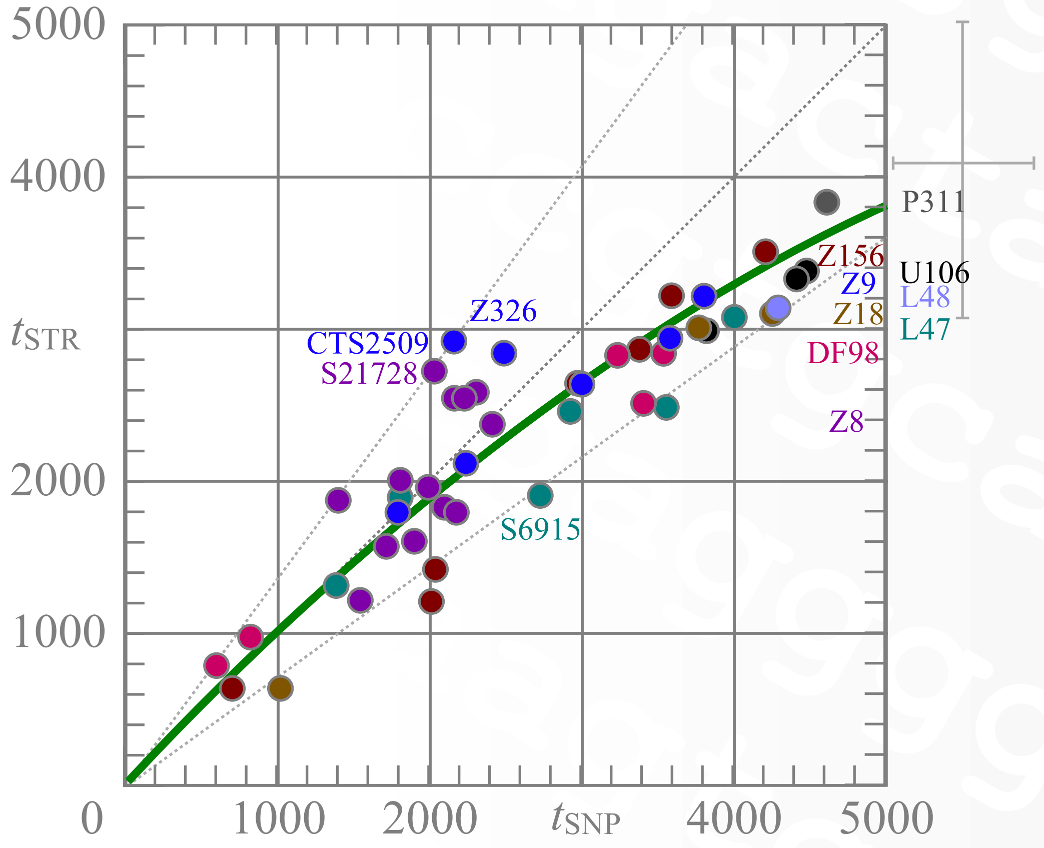
$$t_{\mathrm{STR,corr}} = t_{\mathrm{STR,unc}} \exp(-t_{\mathrm{STR,unc}}/f)$$

where $t_{\mathrm{STR,corr}}$ and $t_{\mathrm{STR,unc}}$ are the uncorrected and corrected ages derived from STRs, respectively, and f is a fitted scaling factor based on calibration to the SNP-derived ages. We fit the following formula:

$$t_{\mathrm{STR,corr}} = t_{\mathrm{STR,unc}}\, f_2 \exp(-t_{\mathrm{STR,unc}}/f_1)$$

with two fitting factors, to allow for uncertainties in the systematic calibration of both ages.

The graph below shows the scaling factors derived for the variance method. The details of this data and the fit can be found in the supplementary spreadsheet (str-ages.xls).



The points are colour-coded following the same clade-based scheme used later in the document. The diagonal gray line represents a 1:1 correlation, and the adjacent lines show the tolerable range of fits based on the typical systematic uncertainty (illustrated on the right). The solid green line shows our best fit. The following fitting parameters are derived:

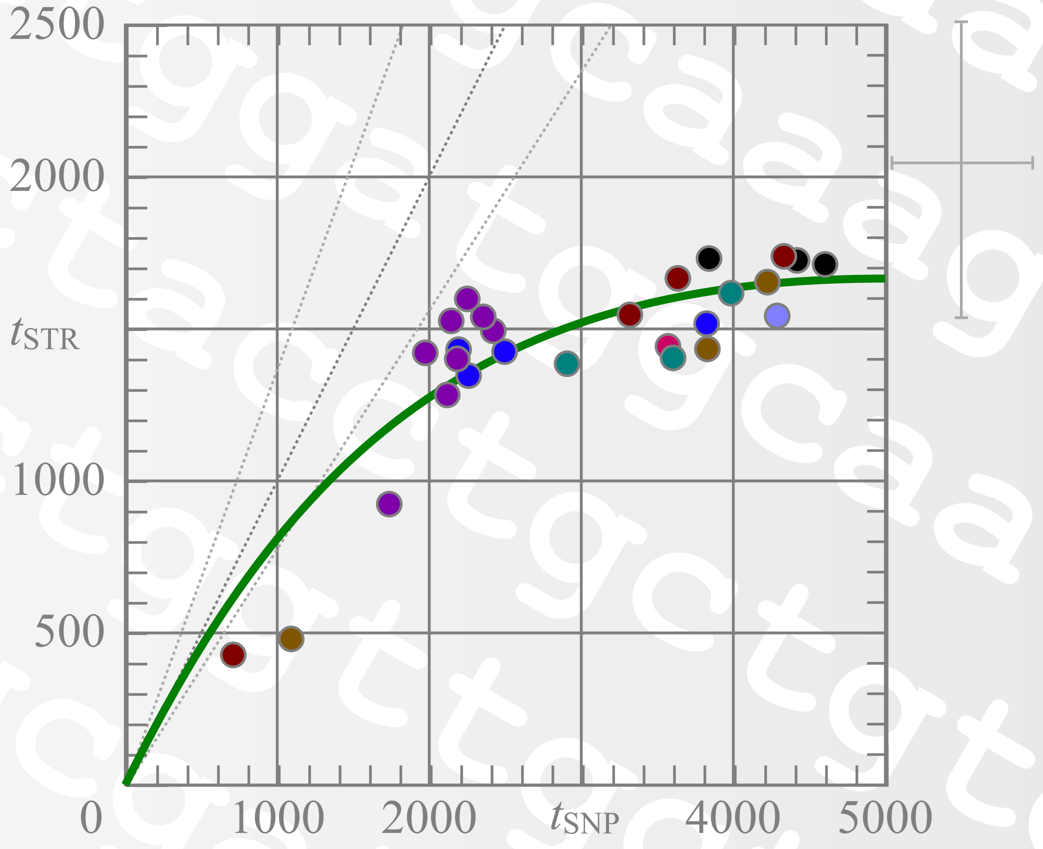
$$f = 18410 \pm 2096,\quad f_1 = 15000 \pm 4114,\quad f_2 = 1.04 \pm 0.07.$$

Corrections become important for this method if the predicted age exceeds about 2500 years old. No statistical difference is determined between the one- and two-parameter fits.

Scatter between the two age estimates is around +/- 300 years (standard deviation). This could be reduced by adopting the same "top-down" constraint to the STR-based ages as is currently applied to the SNP-based ages.

CALIBRATION OF STR TO SNP AGES: INFINITE ALLELES

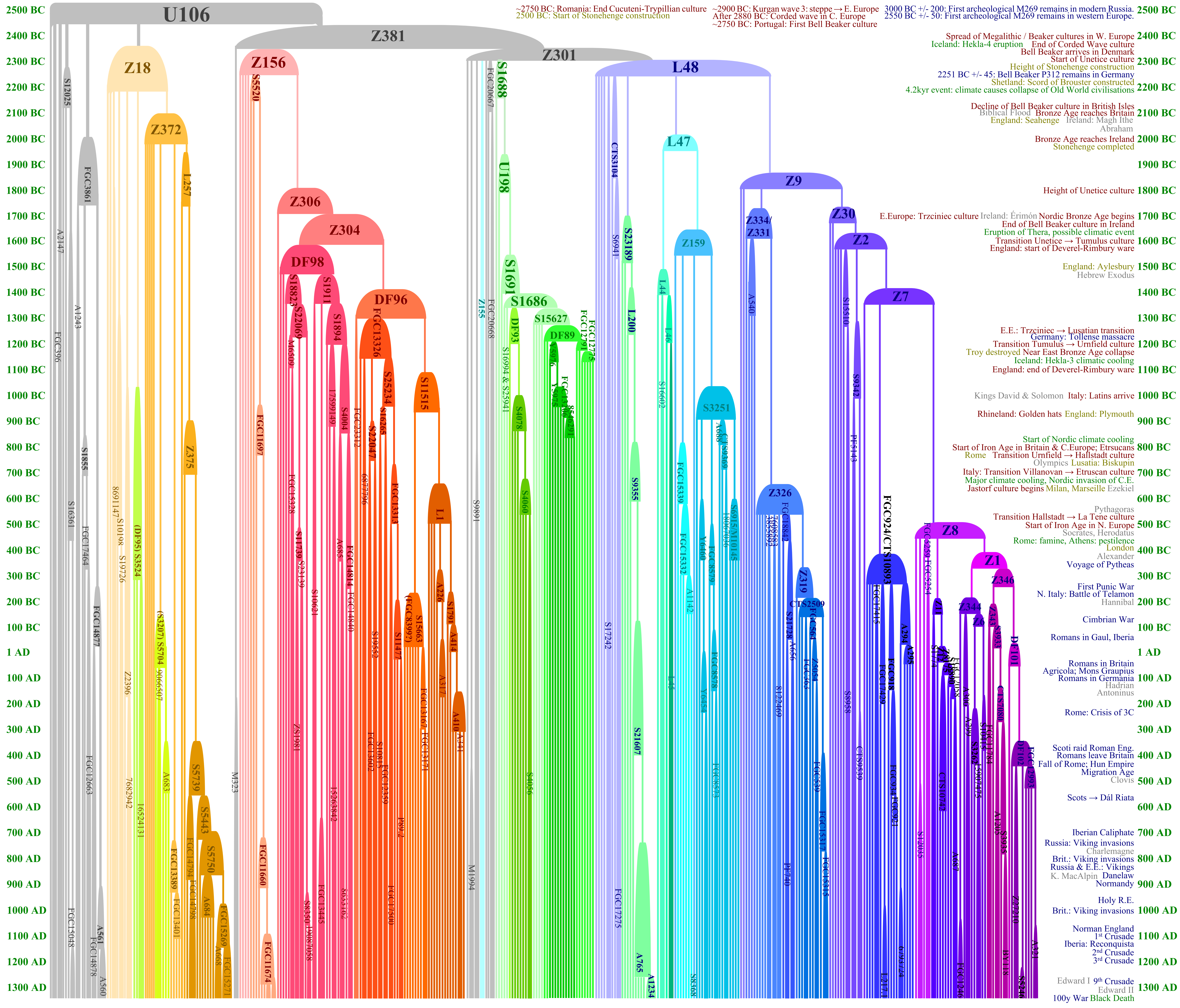
The figure below is similar to the one in the previous panel, except for the infinite alleles method. Note the expanded range on the vertical axis.



The following fitting parameters are derived:

$$f = 4436 \pm 117,\quad f_1 = 4519 \pm 451,\quad f_2 = 0.99 \pm 0.07.$$

Corrections may become important at any age. Again, no significant statistical difference is found between the one- and two-parameter fits. The ages asymptote to a much younger age (around 1700 years). Corrections become important in less than 1000 years, and ages more than about 2000 years cannot be meaningfully corrected.



U106 family tree

Updated: 09 Apr 2015; Dr. Iain McDonald for the U106/S21 group

DESCRIPTION

This phylogenic tree of U106 shows the relationships between the 366 testers with Family Tree DNA BigY results as of 16 Mar 2015, along with the SNP names that define those relationships. Origin dates (to be read at the bottom of the SNP name) are computed using SNP counting.

INTERPRETING DATES

These have been calibrated to existing data using: 133.74 years per SNP. A 95% confidence interval of 119-150 years per SNP has been estimated, hence the absolute ages of each clade are uncertain by +/-12%. For U106 itself, this translates to around +/-560 years. Additional uncertainties due to random sampling of lines become important in smaller lineages. Exact dates should therefore be interpreted very carefully, without placing too much emphasis on particular short-term events.

These dates have been checked against those measured for the subclades L21 and DF27 on the "brother" branch to U106, P312. The U106 date implies an origin for the upstream P311 subclade of around 2760 BC. The date derived from L21 for P311 is 3070 BC. The date derived from DF27 is 2990 BC. These dates are all derived using the same methodology and are thus all affected by the number of years per SNP chosen as a reference value.

A direct comparison comes from archaeological DNA remains. Haplogroup R is not found in Europe before 2600 BC, despite extensive testing of archaeological remains from prior millenia. This provides very strong evidence that the major incursion both R1a and R1b into Europe occurred around this date. R1a and R1b remains are first found roughly simultaneously in south-eastern Germany (Kromsdorf and Ergolding) in the period 2600 - 2500 BC. It is clear from the archaeology that the migration into Europe of the ancestors of U106 happened some time in the period 3000 - 2500 BC.

It is reasonable to assume that such an incursion is associated with a population expansion, leading to the production of many new SNPs. At the origin of this expansion is likely to be P311, which greatly dominates all R1b in Europe.

The timing of the origin of U106 and P311 can logically be linked to archaeolgical horizons during this period, hence the expansion of a series of cultures westwards across Europe during the period 3300 - 2200 BC.

INTERPRETING STRUCTURE

The size of individual SNPs is the product of two factors: their relative size today and the relative frequency at which people test. For example, it is clear that Z381 was ultimately more successful than its brother clade, Z17640. However, populations with dominant populations in the UK and in France (e.g.) will have very different sizes, since very few French men have tested, while many British men have tested.

Much can be guessed from the large gaps and blooms in the phylogenic record. For example, the large gap between L48 and Z9 or Z7 and Z8 could be due to a population crash during this period, or simply the dominant population migrating out of an area where it is well tested. Conversely, the large number of lines stemming from Z2 and rapid succession of SNPs between Z9 and Z7 could indicate rapid population growth, or the migration of a population to a better-sampled area. Interpreting this diagram is therefore best done in the context of a wider geographical analysis, which we present later in this document.

INITIAL COMMENTS

The structure of this chart points to an initial expansion of U106 that propagated the both lines displayed in grey, the top structure of L48 and the U198 precursor, S1688. Random differences in the number of SNPs could mean that Z18, Z372 and L257 formed the tail end of this population expansion, or they could have come at a later date. It is clear, however, that all major branches then suffered a hiatus in population expansion, or a population contraction around this time, as a gap of several centuries in present in each of these major lineages.

The structure of Z18 appears to have become frozen in shortly after its appearance with no new SNPs until Z375, a full millenium later. We can surmise that Z18 migrated to an area where it was not able to participate in any major expansions until the post-Roman Migration Age period.

Z306 shows a complex structure, indicating a steady growth of many of its major lineages. This can probably be linked to a period of relative stability, both geographically and socially, The creation of new SNPs does not appear totally random, with bouts of expansion around 1300 BC (+/- 500 years), 800 BC (+/- 400 years), 400 BC (+/- 400 years), 100 BC (+/- 400 years), 500 AD (+/- 300 years) and 1000 AD (+/- 200 years).

U198 shows a structure evolving later, mainly during the European Bronze Age, indicating a rapid expansion around this time (1300 BC +/- 600 years).

L48 shows a more bursting structure, with short burns containing several SNPs (e.g. Z9, Z30, Z2, Z7) each with many branches. These are typically followed by long hiatus periods (e.g. Z159 to S3251, Z7 to Z8).

The geography of U106 testers

Updated: 26 January 2015
Dr. Iain McDonald
on behalf of the U106/S21 group

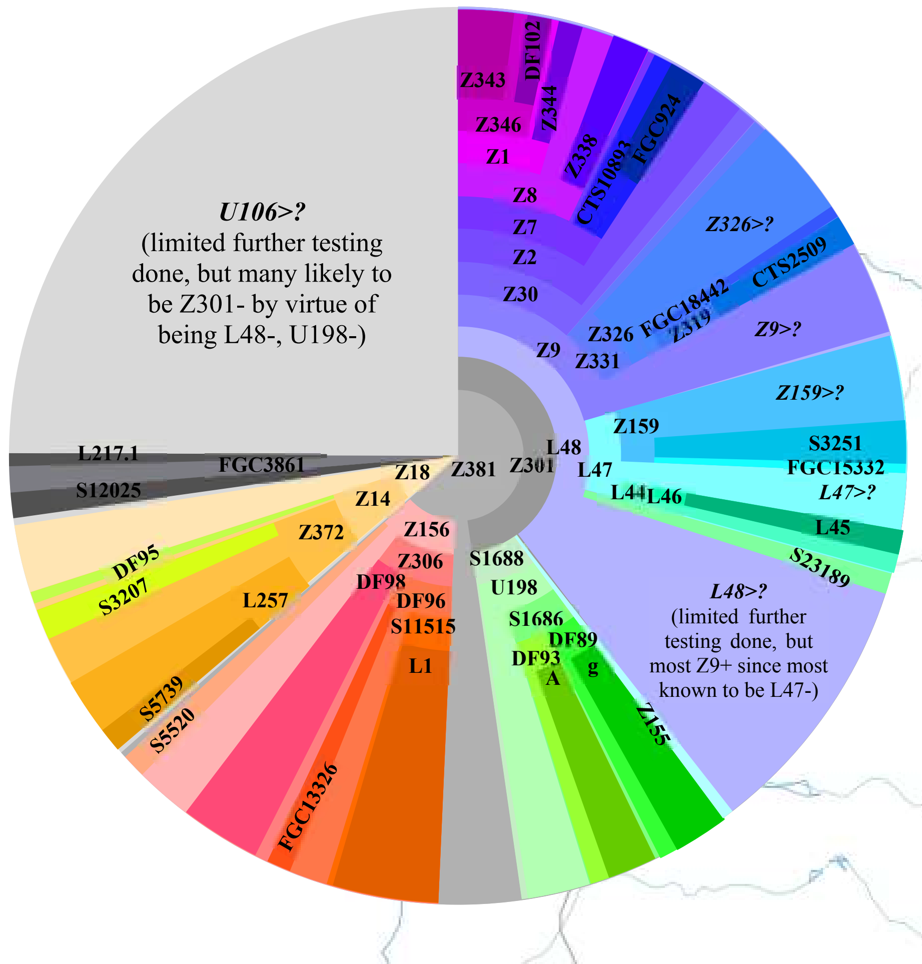
Distribution of all U106 by region and sub-clade

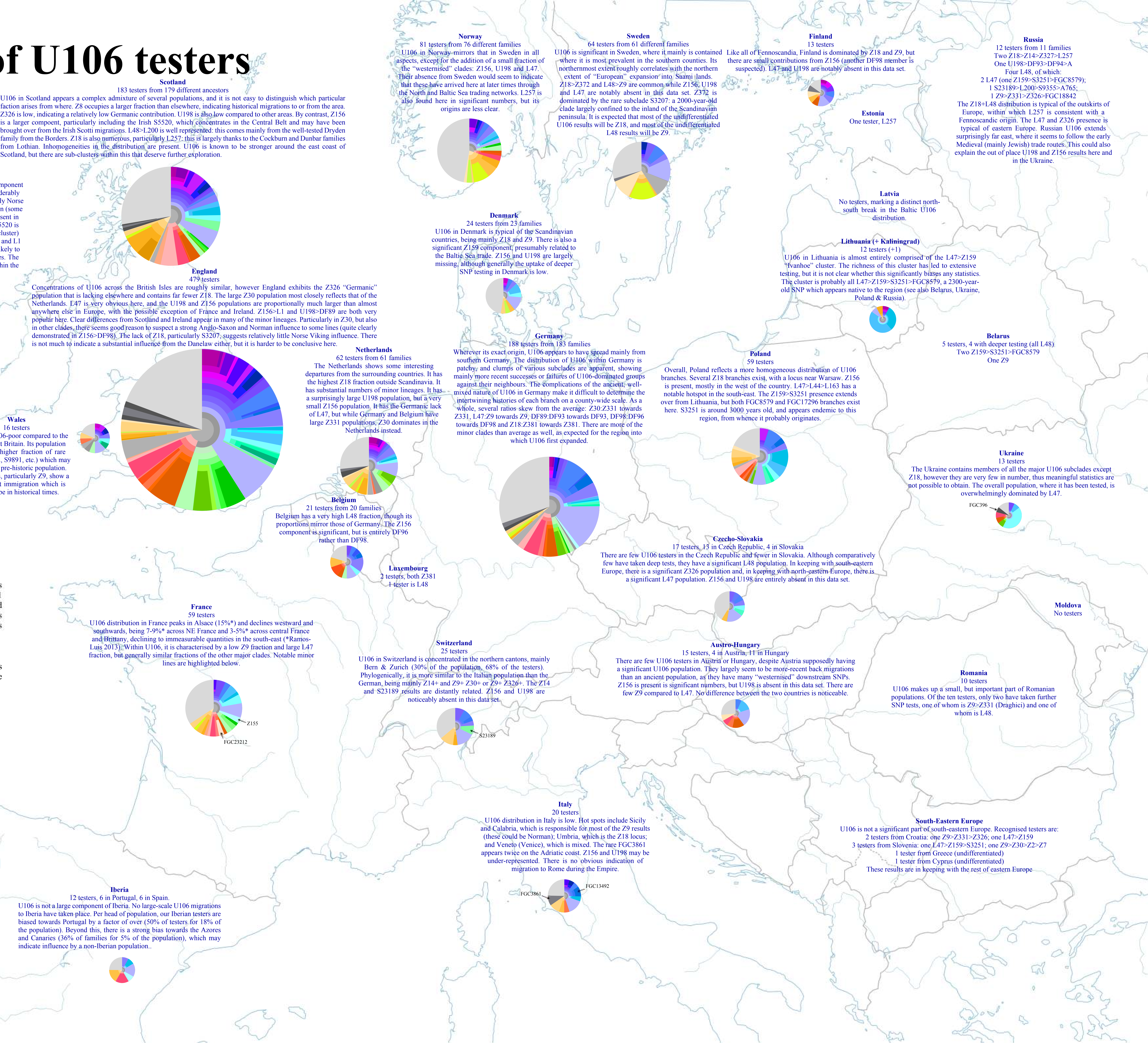
This represents the geographical and phylogenic distribution of 1576 U106+ tests from Family Tree DNA. These include members from the U106, U198 and L1 haplogroup projects, and several geographical projects. Information was collected during December 2014 by several members of the U106 project. Geography is self-reported by the testers. Phylogenic position is based on the last SNP tested, thus some testers may branch into additional, untested subclades.

● = 1 tested person

The proportions of each clade are given for each region. The size of the pie chart is scaled to the size of the tested population. Note that multiple tests from the same family will skew the distributions, and that this has not been accounted for here.

Overall European structure



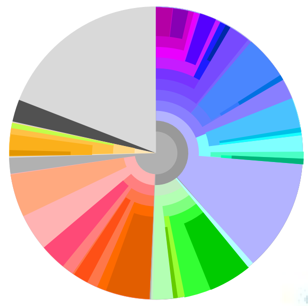


Scotland
183 testers from 179 different ancestors

U106 in Scotland appears a complex admixture of several populations, and it is not easy to distinguish which particular faction arises from where. Z8 occupies a larger fraction than elsewhere, indicating historical migrations to or from the area. Z326 is low, indicating a relatively low Germanic contribution. U198 is also low compared to other areas. By contrast, Z156 is a larger compoent, particularly including the Irish S5520, which concentrates in the Central Belt and may have been brought over from the Irish Scotti migrations. L48>L200 is well represented: this comes mainly from the well-tested Dryden family from the Borders. Z18 is also numerous, particularly L257: this is largely thanks to the Cockburn and Dunbar families from Lothian. Inhomogeneities in the distribution are present. U106 is known to be stronger around the east coast of Scotland, but there are sub-clusters within this that deserve further exploration.

Ireland (Republic of and Northern)
170 testers from 169 families

Although greatly dominated by its brother clade P312, Irish U106 are an important component of the population. Although most branches are represented, the proportions are considerably skewed from the continental average. There are very few Z18, which could be of mostly Norse origin. L48 shows proportions roughly consistent with a north-western European origin (some admixture of the populations bounding the North Sea). Z156 and U198 are both present in large numbers, but they are dominated by particular downstream clades. In Z156, S5520 is very strong, and its large 1000-year-old subclade FGC11660 (the “Mac Maolain” cluster) appears native. Z156>DF96 is also very populous, buoyed by both higher FGC13326 and L1 proportions. Several DF98 here are thought to have an Ulster Scots origin, which is likely to apply to other clades too. In U198, the DF89 (particularly “type g”) clade dominates. The subclades present in Ireland point to most of the U106 population having arrived within the last 2000 years, though as with everywhere there will be exceptions.



England
479 testers

Concentrations of U106 across the British Isles are roughly similar, however England exhibits the Z326 “Germanic” population that is lacking elsewhere and contains far fewer Z18. The large Z30 population most closely reflects that of the Netherlands. L47 is very obvious here, and the U198 and Z156 populations are proportionally much larger than almost anywhere else in Europe, with the possible exception of France and Ireland. Z156>L1 and U198>DF89 are both very popular here. Clear differences from Scotland and Ireland appear in many of the minor lineages. Particularly in Z30, but also in other clades, there seems good reason to suspect a strong Anglo-Saxon and Norman influence to some lines (quite clearly demonstrated in Z156>DF98). The lack of Z18, particularly S3207, suggests relatively little Norse Viking influence. There is not much to indicate a substantial influence from the Danelaw either, but it is harder to be conclusive here.

Wales
16 testers

Wales is U106-poor compared to the rest of Great Britain. Its population includes a higher fraction of rare clades (M232, S9891, etc.) which may be an older, pre-historic population. Other groups, particularly Z9, show a more recent immigration which is likely to be in historical times.

Norway
81 testers from 76 different families

U106 in Norway mirrors that in Sweden in all aspects, except for the addition of a small fraction of the “westernised” clades: Z156, U198 and L47. Their absence from Sweden would seem to indicate that these have arrived here at later times through the North and Baltic Sea trading networks. L257 is also found here in significant numbers, but its origins are less clear.

Sweden
64 testers from 61 different families

U106 is significant in Sweden, where it mainly is contained where it is most prevalent in the southern counties. Its northernmost extent roughly correlates with the northern extent of “European” expansion into Saami lands. Z18>Z372 and L48>Z9 are common while Z156, U198 and L47 are notably absent in this data set. Z372 is dominated by the rare subclade S3207: a 2000-year-old clade largely confined to the inland of the Scandinavian peninsula. It is expected that most of the undifferentiated U106 results will be Z18, and most of the undifferentiated L48 results will be Z9.

Finland
13 testers

Like all of Fennoscandia, Finland is dominated by Z18 and Z9, but there are small contributions from Z156 (another DF98 member is suspected). L47 and U198 are notably absent in this data set.

Russia
12 testers from 11 families
Two Z18>Z14>Z327>L257
One U198>DF93>DF94>A
Four L48, of which:
2 L47 (one Z159>S3251>FGC8579);
1 S23189>L200>S9355>A765;
1 Z9>Z331>Z326>FGC18842

The Z18+L48 distribution is typical of the outskirts of Europe, within which L257 is consistent with a Fennoscandic origin. The L47 and Z326 presence is typical of eastern Europe. Russian U106 extends surprisingly far east, where it seems to follow the early Medieval (mainly Jewish) trade routes. This could also explain the out of place U198 and Z156 results here and in the Ukraine.

Estonia
One tester, L257

Latvia
No testers, marking a distinct north-south break in the Baltic U106 distribution.

Denmark
24 testers from 23 families

U106 in Denmark is typical of the Scandinavian countries, being mainly Z18 and Z9. There is also a significant Z159 component, presumably related to the Baltic Sea trade. Z156 and U198 are largely missing, although generally the uptake of deeper SNP testing in Denmark is low.

Lithuania (+ Kaliningrad)
12 testers (+1)

U106 in Lithuania is almost entirely comprised of the L47>Z159 “Ivanhoe” cluster. The richness of this cluster has led to extensive testing, but it is not clear whether this significantly biases any statistics. The cluster is probably all L47>Z159>S3251>FGC8579, a 2300-year-old SNP which appears native to the region (see also Belarus, Ukraine, Poland & Russia).

Belarus
5 testers, 4 with deeper testing (all L48)
Two Z159>S3251>FGC8579
One Z9

Netherlands
62 testers from 61 families

The Netherlands shows some interesting departures from the surrounding countries. It has the highest Z18 fraction outside Scandinavia. It has substantial numbers of minor lineages. It has a surprisingly large U198 population, but a very small Z156 population. It has the Germanic lack of L47, but while Germany and Belgium have large Z331 populations, Z30 dominates in the Netherlands instead.

Belgium
21 testers from 20 families

Belgium has a very high L48 fraction, though its proportions mirror those of Germany. The Z156 component is significant, but is entirely DF96 rather than DF98.

Luxembourg
2 testers, both Z381
1 tester is L48

Germany
188 testers from 183 families

Wherever its exact origin, U106 appears to have spread mainly from southern Germany. The distribution of U106 within Germany is patchy, and clumps of various subclades are apparent, showing mainly more recent successes or failures of U106-dominated groups against their neighbours. The complications of the ancient, well-mixed nature of U106 in Germany make it difficult to determine the intertwining histories of each branch on a county-wide scale. As a whole, several ratios skew from the average: Z30:Z331 towards Z331, L47:Z9 towards Z9, DF89:DF93 towards DF93, DF98:DF96 towards DF98 and Z18:Z381 towards Z381. There are more of the minor clades than average as well, as expected for the region into which U106 first expanded.

Poland
59 testers

Overall, Poland reflects a more homogeneous distribution of U106 branches. Several Z18 branches exist, with a locus near Warsaw. Z156 is present, mostly in the west of the country. L47>L44>L163 has a notable hotspot in the south-east. The Z159>S3251 presence extends over from Lithuania, but both FGC8579 and FGC17296 branches exist here. S3251 is around 3000 years old, and appears endemic to this region, from whence it probably originates.

Ukraine
13 testers

The Ukraine contains members of all the major U106 subclades except Z18, however they are very few in number, thus meaningful statistics are not possible to obtain. The overall population, where it has been tested, is overwhelmingly dominated by L47.

Czecho-Slovakia
17 testers, 13 in Czech Republic, 4 in Slovakia

There are few U106 testers in the Czech Republic and fewer in Slovakia. Although comparatively few have taken deep tests, they have a significant L48 population. In keeping with south-eastern Europe, there is a significant Z326 population and, in keeping with north-eastern Europe, there is a significant L47 population. Z156 and U198 are entirely absent in this data set.

Moldova
No testers

France
59 testers

U106 distribution in France peaks in Alsace (15%*) and declines westward and southwards, being 7-9%* across NE France and 3-5%* across central France and Brittany, declining to immeasurable quantities in the south-east (*Ramos-Luis 2013). Within U106, it is characterised by a low Z9 fraction and large L47 fraction, but generally similar fractions of the other major clades. Notable minor lines are highlighted below.

Switzerland
25 testers

U106 in Switzerland is concentrated in the northern cantons, mainly Bern & Zurich (30% of the population, 68% of the testers). Phylogenically, it is more similar to the Italian population than the German, being mainly Z14+ and Z9+ Z30+ or Z9+ Z326+. The Z14 and S23189 results are distantly related. Z156 and U198 are noticeably absent in this data set.

Austro-Hungary
15 testers, 4 in Austria, 11 in Hungary

There are few U106 testers in Austria or Hungary, despite Austria supposedly having a significant U106 population. They largely seem to be more-recent back migrations than an ancient population, as they have many “westernised” downstream SNPs. Z156 is present is significant numbers, but U198 is absent in this data set. There are few Z9 compared to L47. No difference between the two countries is noticeable.

Romania
10 testers

U106 makes up a small, but important part of Romanian populations. Of the ten testers, only two have taken further SNP tests, one of whom is Z9>Z331 (Draghici) and one of whom is L48.

Italy
20 testers

U106 distribution in Italy is low. Hot spots include Sicily and Calabria, which is responsible for most of the Z9 results (these could be Norman); Umbria, which is the Z18 locus; and Veneto (Venice), which is mixed. The rare FGC3861 appears twice on the Adriatic coast. Z156 and U198 may be under-represented. There is no obvious indication of migration to Rome during the Empire.

South-Eastern Europe

U106 is not a significant part of south-eastern Europe. Recognised testers are:
2 testers from Croatia: one Z9>Z331>Z326; one L47>Z159
3 testers from Slovenia: one L47>Z159>S3251; one Z9>Z30>Z2>Z7
1 tester from Greece (undifferentiated)
1 tester from Cyprus (undifferentiated)
These results are in keeping with the rest of eastern Europe

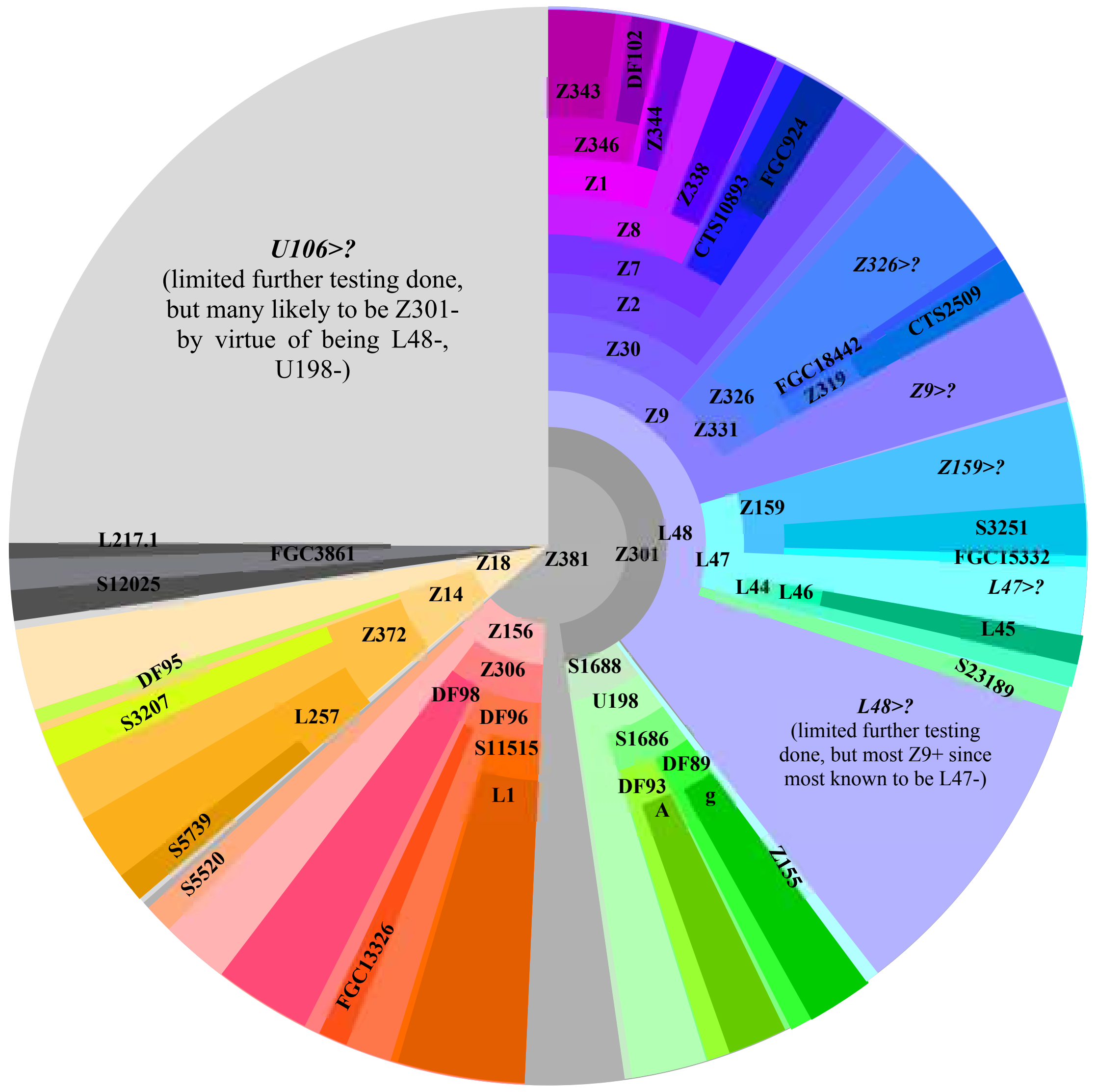
Iberia
12 testers, 6 in Portugal, 6 in Spain.

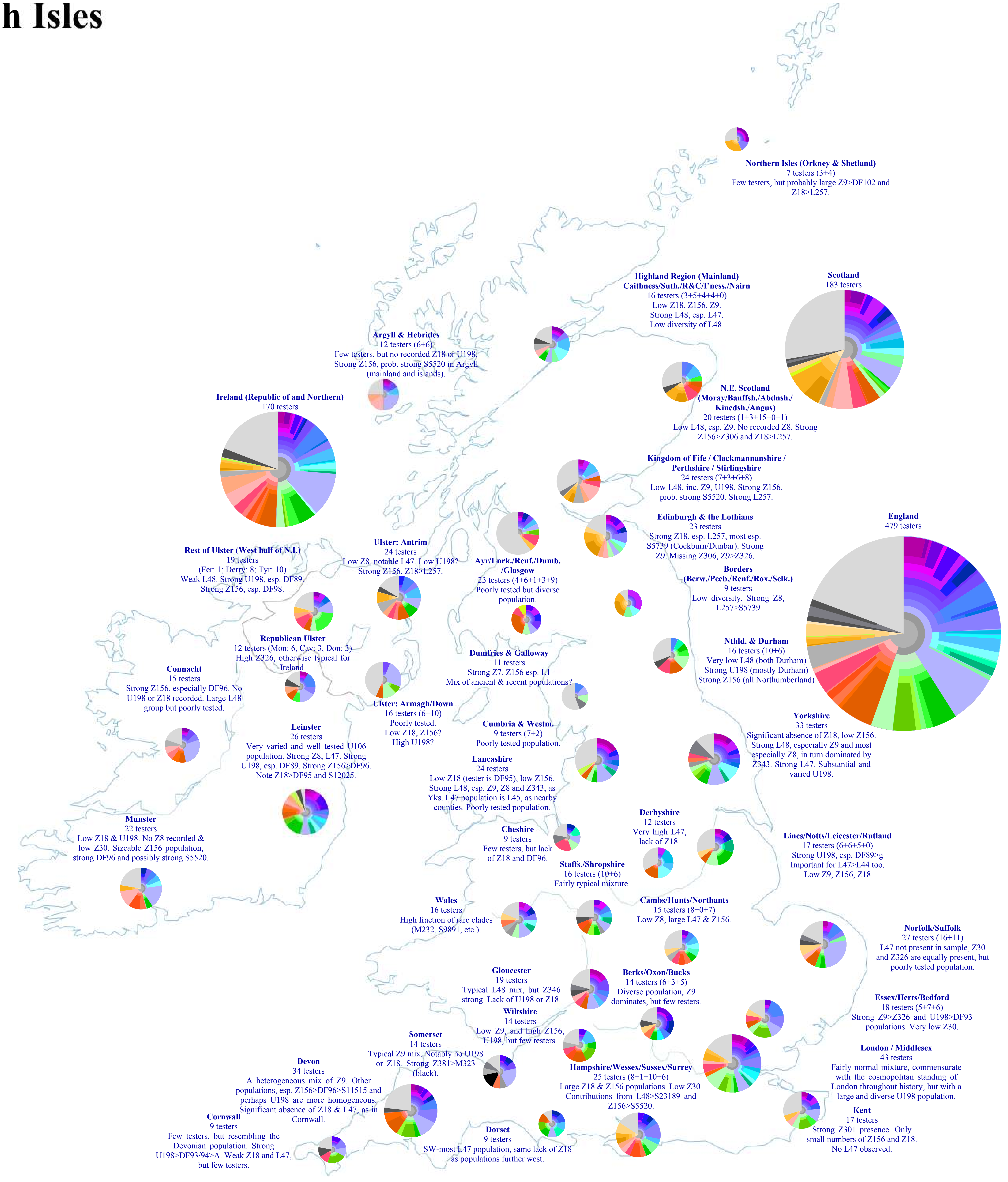
U106 is not a large component of Iberia. No large-scale U106 migrations to Iberia have taken place. Per head of population, our Iberian testers are biased towards Portugal by a factor of over (50% of testers for 18% of the population). Beyond this, there is a strong bias towards the Azores and Canaries (36% of families for 5% of the population), which may indicate influence by a non-Iberian population..

The geography of U106 testers in the British Isles

Updated: 25 March 2015
Dr. Iain McDonald
on behalf of the U106/S21 group

All of Europe





Northern Isles (Orkney & Shetland)
7 testers (3+4)
Few testers, but probably large Z9>DF102 and Z18>L257.

Highland Region (Mainland) Caithness/Suth./R&C/I'ness./Nairn
16 testers (3+5+4+4+0)
Low Z18, Z156, Z9.
Strong L48, esp. L47.
Low diversity of L48.

Scotland
183 testers

Argyll & Hebrides
12 testers (6+6)
Few testers, but no recorded Z18 or U198. Strong Z156, prob. strong S5520 in Argyll (mainland and islands).

N.E. Scotland (Moray/Banffsh./Abdnsh./ Kincdsh./Angus)
20 testers (1+3+15+0+1)
Low L48, esp. Z9. No recorded Z8. Strong Z156>Z306 and Z18>L257.

Ireland (Republic of and Northern)
170 testers

Kingdom of Fife / Clackmannanshire / Perthshire / Stirlingshire
24 testers (7+3+6+8)
Low L48, inc. Z9, U198. Strong Z156, prob. strong S5520. Strong L257.

Edinburgh & the Lothians
23 testers
Strong Z18, esp. L257, most esp. S5739 (Cockburn/Dunbar). Strong Z9. Missing Z306, Z9>Z326.

England
479 testers

Rest of Ulster (West half of N.I.)
19 testers
(Fer: 1; Derry: 8; Tyr: 10)
Weak L48. Strong U198, esp. DF89. Strong Z156, esp. DF98.

Ulster: Antrim
24 testers
Low Z8, notable L47. Low U198? Strong Z156, Z18>L257.

Ayr/Lnrk./Renf./Dumb. /Glasgow
23 testers (4+6+1+3+9)
Poorly tested but diverse population.

Borders (Berw./Peeb./Renf./Rox./Selk.)
9 testers
Low diversity. Strong Z8, L257>S5739

Republican Ulster
12 testers (Mon: 6, Cav: 3, Don: 3)
High Z326, otherwise typical for Ireland.

Nthld. & Durham
16 testers (10+6)
Very low L48 (both Durham)
Strong U198 (mostly Durham)
Strong Z156 (all Northumberland)

Dumfries & Galloway
11 testers
Strong Z7, Z156 esp. L1
Mix of ancient & recent populations?

Connacht
15 testers
Strong Z156, especially DF96. No U198 or Z18 recorded. Large L48 group but poorly tested.

Ulster: Armagh/Down
16 testers (6+10)
Poorly tested.
Low Z18, Z156?
High U198?

Yorkshire
33 testers
Significant absence of Z18, low Z156. Strong L48, especially Z9 and most especially Z8, in turn dominated by Z343. Strong L47. Substantial and varied U198.

Leinster
26 testers
Very varied and well tested U106 population. Strong Z8, L47. Strong U198, esp. DF89. Strong Z156>DF96. Note Z18>DF95 and S12025.

Cumbria & Westm.
9 testers (7+2)
Poorly tested population.

Lancashire
24 testers
Low Z18 (tester is DF95), low Z156. Strong L48, esp. Z9, Z8 and Z343, as Yks. L47 population is L45, as nearby counties. Poorly tested population.

Munster
22 testers
Low Z18 & U198. No Z8 recorded & low Z30. Sizeable Z156 population, strong DF96 and possibly strong S5520.

Derbyshire
12 testers
Very high L47, lack of Z18.

Cheshire
9 testers
Few testers, but lack of Z18 and DF96.

Lincs/Notts/Leicester/Rutland
17 testers (6+6+5+0)
Strong U198, esp. DF89>g
Important for L47>L44 too.
Low Z9, Z156, Z18

Staffs./Shropshire
16 testers (10+6)
Fairly typical mixture.

Wales
16 testers
High fraction of rare clades (M232, S9891, etc.).

Cambs/Hunts/Northants
15 testers (8+0+7)
Low Z8, large L47 & Z156.

Norfolk/Suffolk
27 testers (16+11)
L47 not present in sample, Z30 and Z326 are equally present, but poorly tested population.

Gloucester
19 testers
Typical L48 mix, but Z346 strong. Lack of U198 or Z18.

Berks/Oxon/Bucks
14 testers (6+3+5)
Diverse population, Z9 dominates, but few testers.

Essex/Herts/Bedford
18 testers (5+7+6)
Strong Z9>Z326 and U198>DF93 populations. Very low Z30.

Wiltshire
14 testers
Low Z9, and high Z156, U198, but few testers.

Somerset
14 testers
Typical Z9 mix. Notably no U198 or Z18. Strong Z381>M323 (black).

London / Middlesex
43 testers
Fairly normal mixture, commensurate with the cosmopolitan standing of London throughout history, but with a large and diverse U198 population.

Devon
34 testers
A heterogeneous mix of Z9. Other populations, esp. Z156>DF96>S11515 and perhaps U198 are more homogeneous. Significant absence of Z18 & L47, as in Cornwall.

Hampshire/Wessex/Sussex/Surrey
25 testers (8+1+10+6)
Large Z18 & Z156 populations. Low Z30. Contributions from L48>S23189 and Z156>S5520.

Cornwall
9 testers
Few testers, but resembling the Devonian population. Strong U198>DF93/94>A. Weak Z18 and L47, but few testers.

Dorset
9 testers
SW-most L47 population, same lack of Z18 as populations further west.

Kent
17 testers
Strong Z301 presence. Only small numbers of Z156 and Z18.
No L47 observed.

● = 1 tested person

The proportions of each clade are given for each region. The size of the pie chart is scaled to the size of the tested population. Note that multiple tests from the same family will skew the distributions, and that this has not been accounted for here.

Migrations

METHODS TO IDENTIFY MIGRATION PATHS

There are three primary ways to work out how and when a clade spread.
(1) Look at the current distribution of people. This tells you something about who is related to who, but not when and why.
(2) Look at archaeological DNA results. As already discussed, this is very good at determining upper limits to when clades formed. It can also tell you something about the earliest phases of a population's presence in an area. However, these are only glimpses: snapshots into a forgotten world, and very few and far between. There's only so much information they can give on particular time periods and particular migrations, unless they happen to be very large. Nevertheless, this is the most effective method for older populations.
(3) Look at the ages of MRCAs from different countries. This is perhaps the most powerful tool for more recent populations, but perhaps also the most difficult to obtain good data from. We will discuss this later.

CURRENT DISTRIBUTION

The current relative distribution of U106 and its subclades can be found on the following pages. These come from a compilation of projects at Family Tree DNA, not least the U106 project itself. They are therefore biased by the content of those projects.

Some projects are more active than others at recruiting members,.Some are more active in getting members to test to more-recent SNPs. Some families have DNA tested many members, some only one, despite being of the same size in the present-day population. These fractions should not be taken as absolute proportions, but as guides or indicators for further work.

ARCHAEOLOGICAL Y-DNA

Following the maps of current distribution are several maps showing the archaeological DNA results up to 2000 BC, shortly after U106 formed. These are broken up by period to highlight the differences between them.

From these, it can clearly be seen that haplogroup R was essentially absent from Europe until some time shortly before 2600 BC. It was, however, present in modern Russia, and this has been used to indicate a rapid spread of both R1a and R1b into Europe, during the period circa 3300 BC to 2500 BC (see Haak et al. 2015). This has been associated with the archaeological Kurgan and Yamnaya cultures, and can possibly be credited with bringing Indo-European language and culture to Europe.

THE ROUTE INTO EUROPE

The route our ancestors took into Europe is not precisely known. From the ancient DNA record, R1a and R1b both seem to appear "overnight" sometime during the early third millenium BC.

The exact origin of these people is not known. They may have come from as far south as the southern Caucasus, or as far north as the tundra-covered northern Ural mountains. What we do know is that they somehow spawned the Yamnaya and other cultures who lived north of the Black Sea during the last fourth millenium BC.

It is thought that R1a, at least, helped create the Corded Ware cultural horizon in north-eastern Europe. It is not clear whether or not R1b came with them, due to the relatively smaller number of R1b DNA results during this crucial period.

There are two likely routes of R1b into Europe. The first is to the north of the Carpathian Mountains, through relatively flat plains of modern-day Poland. This follows the Corded Ware culture, and there are some slight preferences for this route from ancient DNA.

The second is down the Black Sea coast and up the River Danube. This is the preferred route from analysing the geography of M269+ U106- P312- Y-DNA results using a "minimal spanning tree"-like method.

A third route, arriving from modern-day Turkey with the advent of farming at the beginning of the Neolithic, appears ruled out by the dates obtained from our results, and the dates and places obtained from ancient DNA.

Distinguishing between the two remaining routes cannot yet be done with confidence. It will need better constraints – either in time or place – from archaeological DNA.

LOCATION OF THE FIRST U106

It is clear that the first major place of U106 settlement was in modern-day Germany, whether it was on the border with Poland, or with Austria, or as far west as the Rhine valley. The exact location depends on the migration pathway into Germany, and the exact time that U106 formed relative to the migration westwards.

Either way, our earliest U106 ancestors were very probably German. U106 has often thought to be the `Germanic' cousin of the `Celtic' P312. In fact, these are misnomers, as both U106 and P312 predate these cultures by over 1000 years. More properly, early U106 probably formed much of the Bell Beaker cultures of central Europe, and later the western half of the Unetice culture.

The earliest (and so far only) ancient U106 burial is dated to between 2275 and 2032 BC, and comes from the Nordic Bronze Age culture of southern Sweden (Lilla Beddinge), rather than Germany. Although likely from several centuries after the formation of U106, this indicates that U106 spread quite quickly and effectively to these areas. Sadly, we do not currently know of any descendents of this particular branch of U106, which may have died out.

Later pages in this section show this formation and dispersion of U106 graphically.

FURTHER ANCIENT DNA

This section is left blank for further DNA results as they arrive.

6500-4000 BC: the situation in Europe

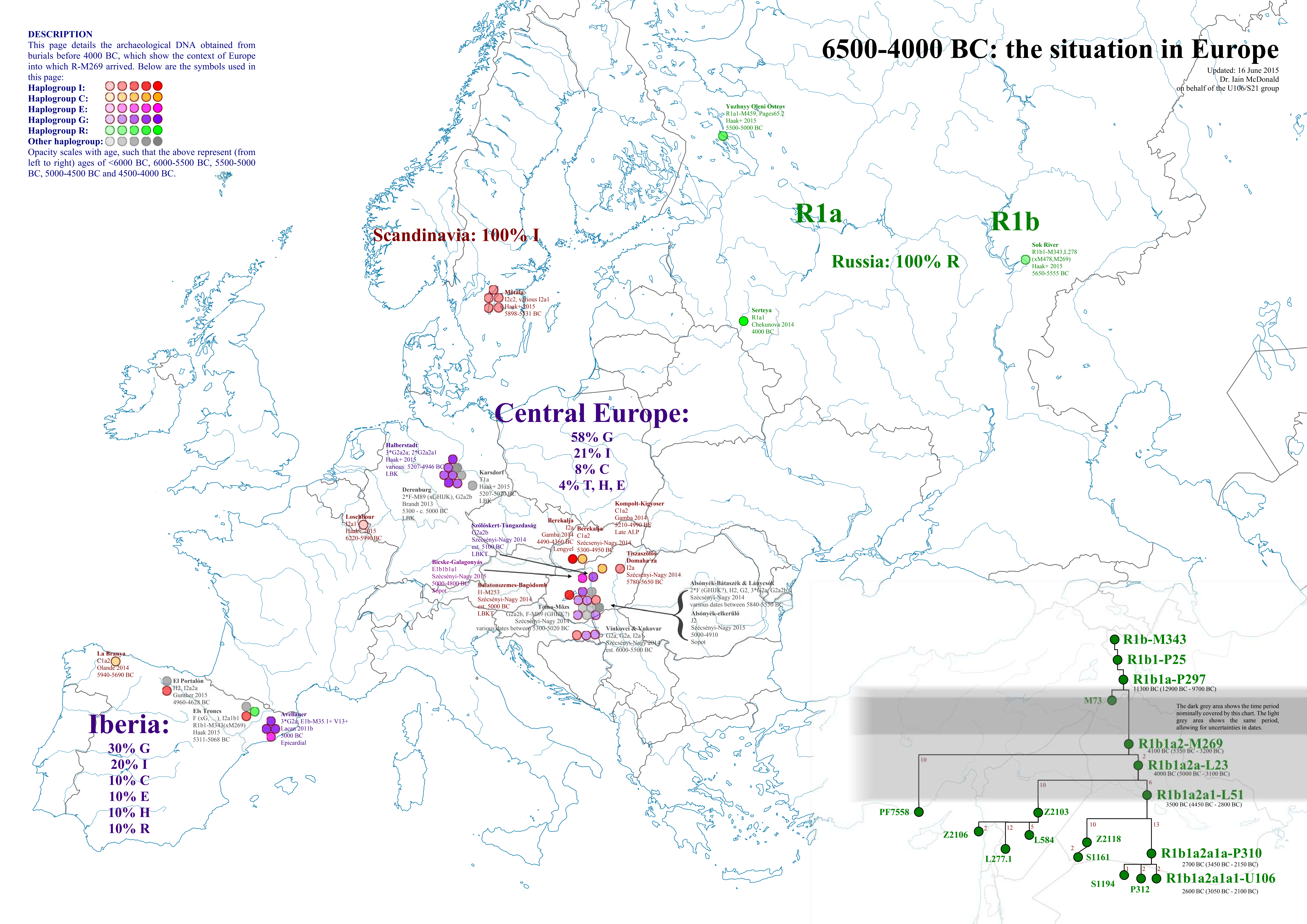
Updated: 16 June 2015
Dr. Iain McDonald
on behalf of the U106/S21 group

DESCRIPTION

This page details the archaeological DNA obtained from burials before 4000 BC, which show the context of Europe into which R-M269 arrived. Below are the symbols used in this page:

Haplogroup I: ● ● ● ● ●
Haplogroup C: ● ● ● ● ●
Haplogroup E: ● ● ● ● ●
Haplogroup G: ● ● ● ● ●
Haplogroup R: ● ● ● ● ●
Other haplogroup: ● ● ● ● ●

Opacity scales with age, such that the above represent (from left to right) ages of <6000 BC, 6000-5500 BC, 5500-5000 BC, 5000-4500 BC and 4500-4000 BC.



The dark grey area shows the time period nominally covered by this chart. The light grey area shows the same period, allowing for uncertainties in dates.

4000-3000 BC: the rise of M269

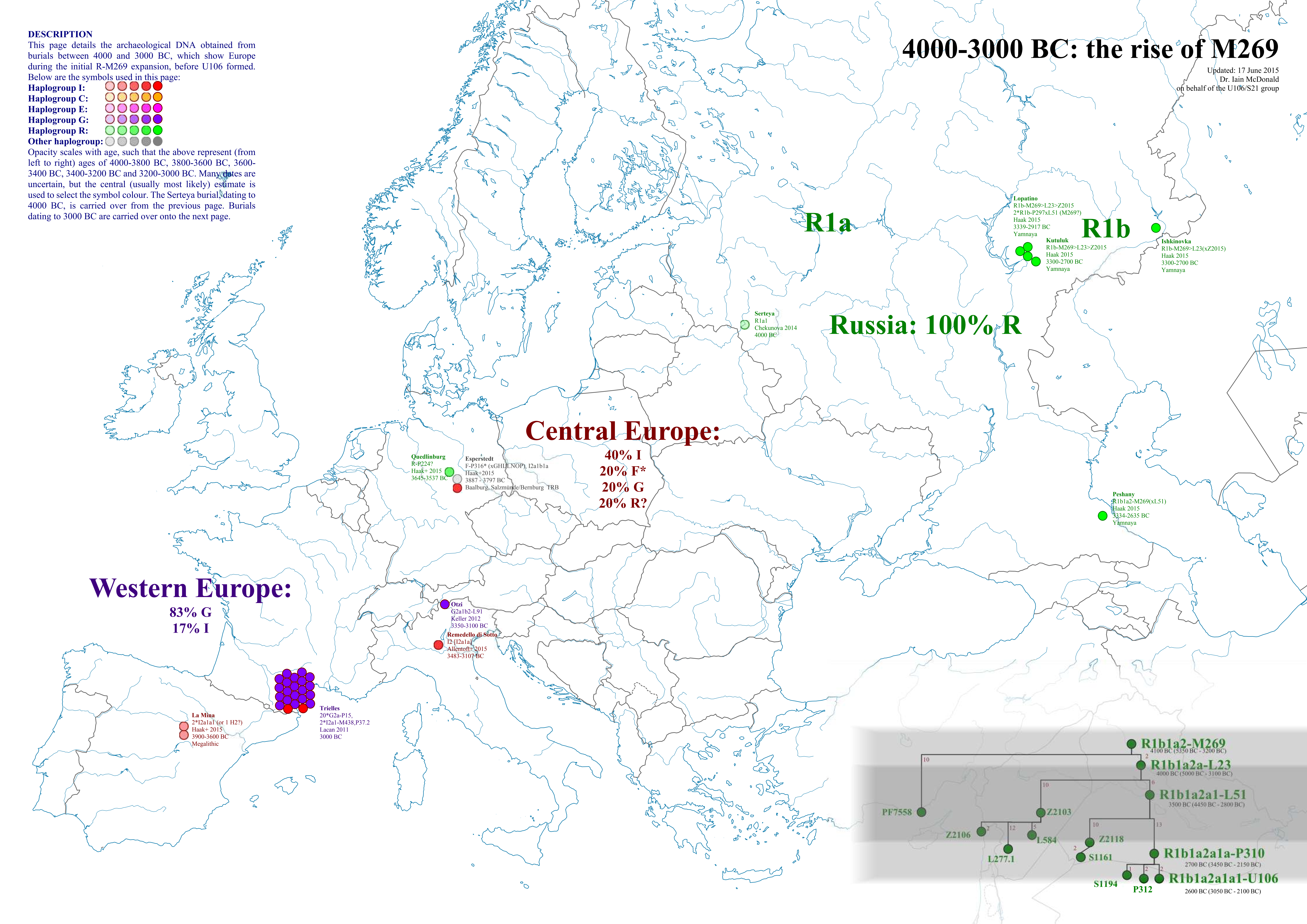
Updated: 17 June 2015
Dr. Iain McDonald
on behalf of the U106/S21 group

DESCRIPTION

This page details the archaeological DNA obtained from burials between 4000 and 3000 BC, which show Europe during the initial R-M269 expansion, before U106 formed. Below are the symbols used in this page:

Haplogroup I: ●●●●●
Haplogroup C: ●●●●●
Haplogroup E: ●●●●●
Haplogroup G: ●●●●●
Haplogroup R: ●●●●●
Other haplogroup: ●●●●●

Opacity scales with age, such that the above represent (from left to right) ages of 4000-3800 BC, 3800-3600 BC, 3600-3400 BC, 3400-3200 BC and 3200-3000 BC. Many dates are uncertain, but the central (usually most likely) estimate is used to select the symbol colour. The Serteya burial, dating to 4000 BC, is carried over from the previous page. Burials dating to 3000 BC are carried over onto the next page.



Lopatino
R1b-M269>L23>Z2015
2*R1b-P297xL51 (M269?)
Haak 2015
3339-2917 BC
Yamnaya

Kutuluk
R1b-M269>L23>Z2015
Haak 2015
3300-2700 BC
Yamnaya

Ishkinovka
R1b-M269>L23(xZ2015)
Haak 2015
3300-2700 BC
Yamnaya

Serteya
R1a1
Chekunova 2014
4000 BC

Peshany
R1b1a2-M269(xL51)
Haak 2015
3334-2635 BC
Yamnaya

Quedlinburg
R-P224?
Haak+ 2015
3645-3537 BC

Esperstedt
F-P316* (xGHIJLNOP), I2a1b1a
Haak+2015
3887 - 3797 BC
Baalburg, Salzmünde/Bernburg TRB

Otzi
G2a1b2-L91
Keller 2012
3350-3100 BC

Remedello di Sotto
I2 [I2a1a]
Allentoft+ 2015
3483-3107 BC

Trielles
20*G2a-P15,
2*I2a1-M438,P37.2
Lacan 2011
3000 BC

La Mina
2*I2a1a1 (or 1 H2?)
Haak+ 2015
3900-3600 BC
Megalithic

R1b1a2-M269
4100 BC (5350 BC - 3200 BC)
R1b1a2a-L23
4000 BC (5000 BC - 3100 BC)
R1b1a2a1-L51
3500 BC (4450 BC - 2800 BC)
PF7558
Z2103
Z2106
L277.1
L584
Z2118
S1161
R1b1a2a1a-P310
2700 BC (3450 BC - 2150 BC)
S1194
P312
R1b1a2a1a1-U106
2600 BC (3050 BC - 2100 BC)
10
2
10
6
2
12
5
10
13
2
1
2
2

3000-2500 BC: M269 invades Europe

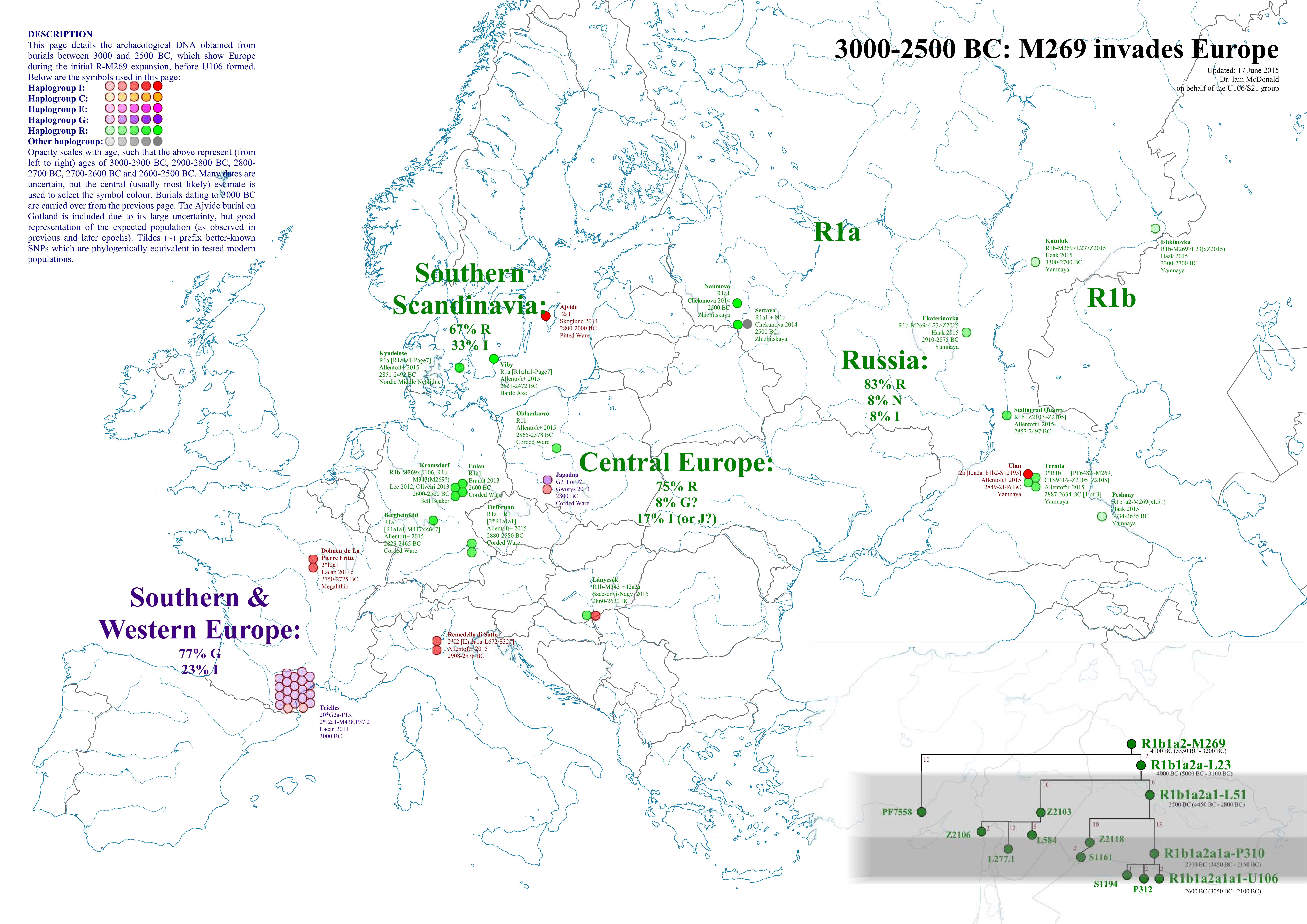
Updated: 17 June 2015
Dr. Iain McDonald
on behalf of the U106/S21 group

DESCRIPTION

This page details the archaeological DNA obtained from burials between 3000 and 2500 BC, which show Europe during the initial R-M269 expansion, before U106 formed. Below are the symbols used in this page:

Haplogroup I: ● ● ● ● ●
Haplogroup C: ● ● ● ● ●
Haplogroup E: ● ● ● ● ●
Haplogroup G: ● ● ● ● ●
Haplogroup R: ● ● ● ● ●
Other haplogroup: ● ● ● ● ●

Opacity scales with age, such that the above represent (from left to right) ages of 3000-2900 BC, 2900-2800 BC, 2800-2700 BC, 2700-2600 BC and 2600-2500 BC. Many dates are uncertain, but the central (usually most likely) estimate is used to select the symbol colour. Burials dating to 3000 BC are carried over from the previous page. The Ajvide burial on Gotland is included due to its large uncertainty, but good representation of the expected population (as observed in previous and later epochs). Tildes (~) prefix better-known SNPs which are phylogenically equivalent in tested modern populations.



2500-2000 BC: early growth of U106

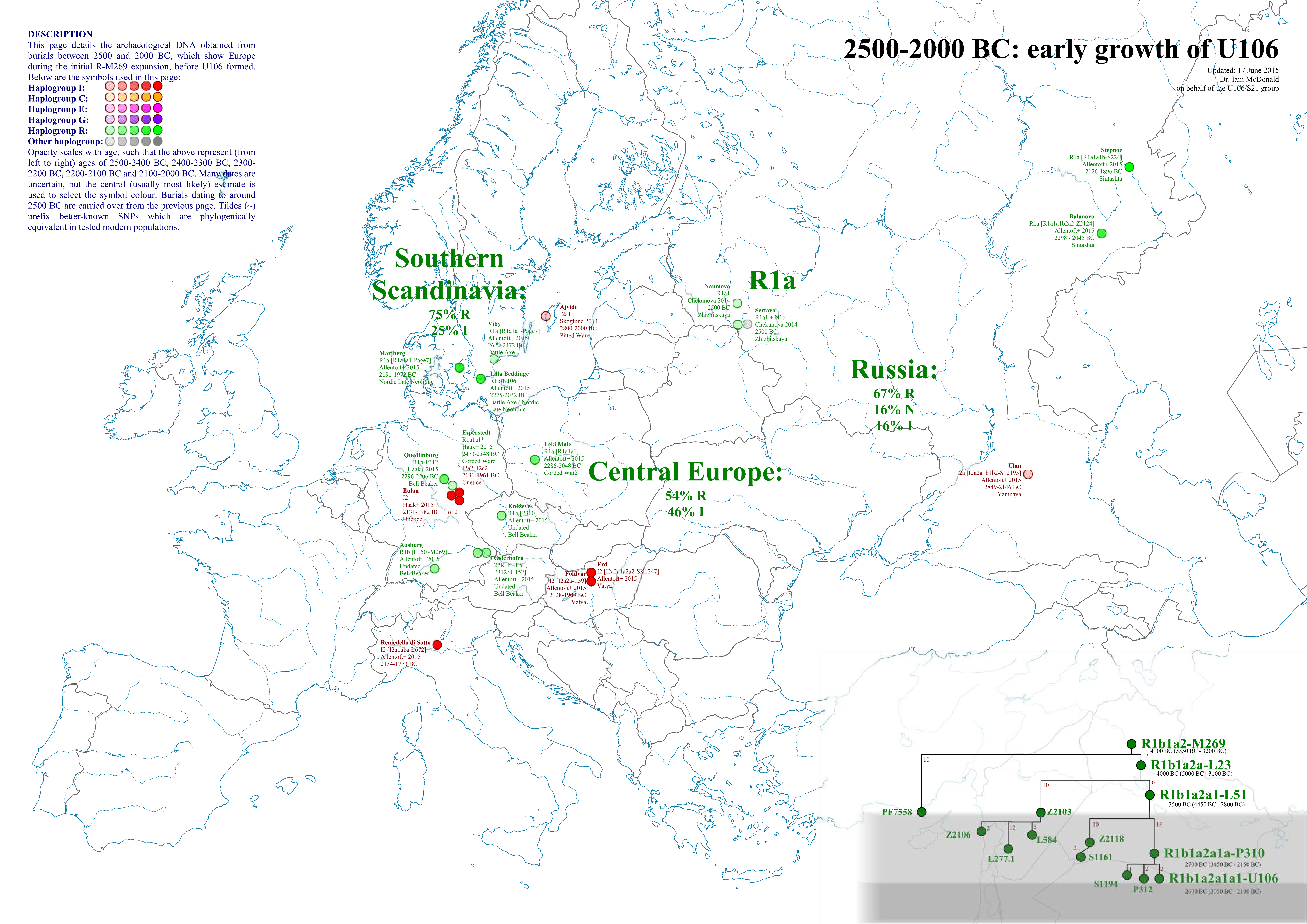
Updated: 17 June 2015
Dr. Iain McDonald
on behalf of the U106/S21 group

DESCRIPTION

This page details the archaeological DNA obtained from burials between 2500 and 2000 BC, which show Europe during the initial R-M269 expansion, before U106 formed. Below are the symbols used in this page:

Haplogroup I: ● ● ● ● ●
Haplogroup C: ● ● ● ● ●
Haplogroup E: ● ● ● ● ●
Haplogroup G: ● ● ● ● ●
Haplogroup R: ● ● ● ● ●
Other haplogroup: ● ● ● ● ●

Opacity scales with age, such that the above represent (from left to right) ages of 2500-2400 BC, 2400-2300 BC, 2300-2200 BC, 2200-2100 BC and 2100-2000 BC. Many dates are uncertain, but the central (usually most likely) estimate is used to select the symbol colour. Burials dating to around 2500 BC are carried over from the previous page. Tildes (~) prefix better-known SNPs which are phylogenically equivalent in tested modern populations.

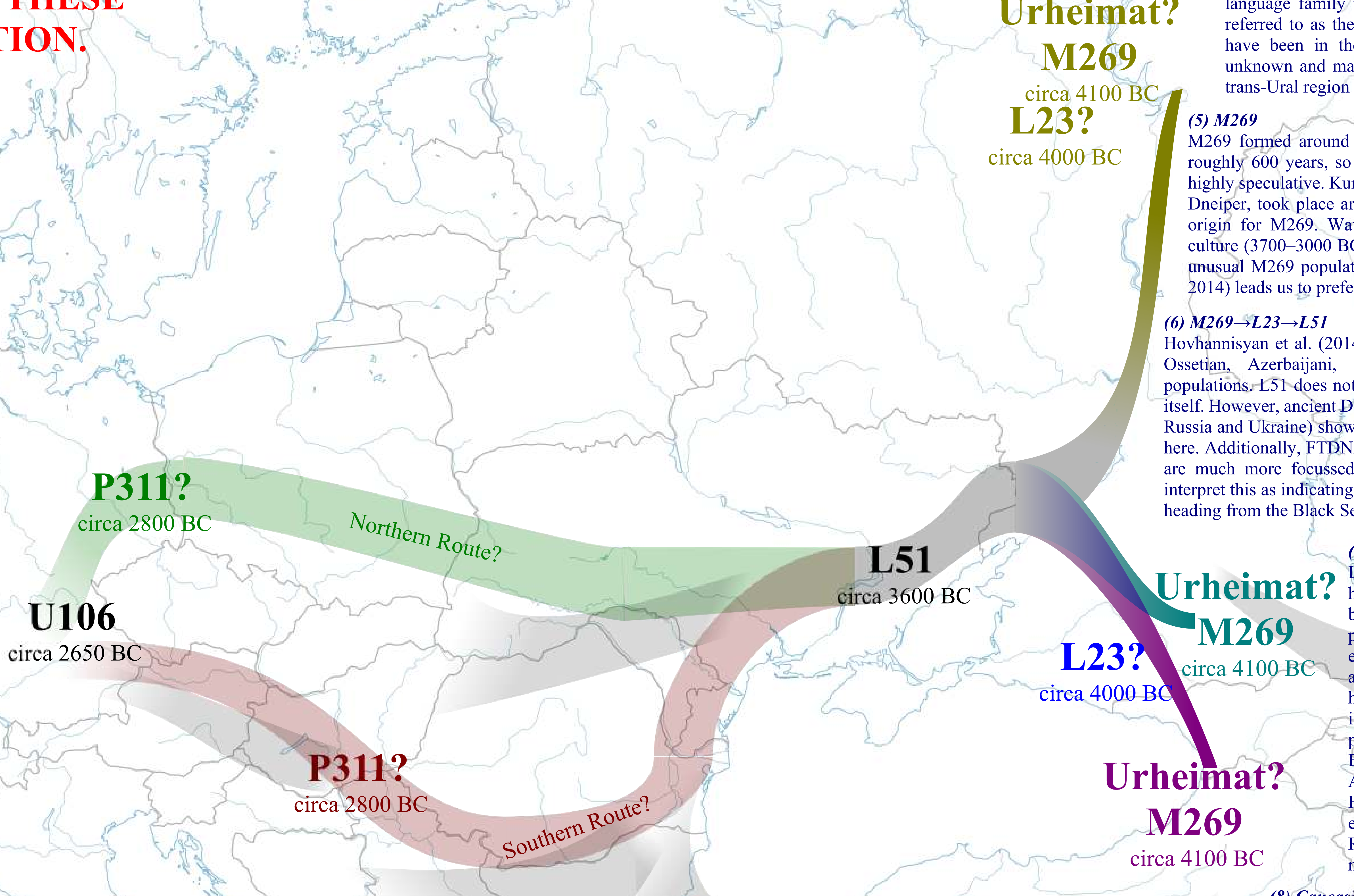


WARNING!

THESE MAPS CONTAIN MANY INFERENCES THAT CANNOT CURRENTLY BE PROVED WITH ANY SCIENTIFIC RIGOUR, BUT WHICH ARE NECESSARY TO PRESENT A COHERENT PICTURE. THEY WILL NOT BE A TRUE REPRESENTATION OF HISTORICAL MIGRATIONS. THESE MAPS ONLY REPRESENT A STARTING HYPOTHESIS FOR CONTINUING WORK. DO NOT TAKE THESE RESULTS AS BEING THE ONLY VIABLE OPTION.

The origins of U106: 3800 BC to 2650 BC

Updated: 20 September 2015
Dr. Iain McDonald
on behalf of the U106/S21 group



(1) Arrival into Europe

The origin of U106 can now be placed around 2650 BC. There is roughly a 2-in-3 chance of it being within 250 years of this date. Ancient DNA shows few haplogroup R men in Europe before about 3000 BC. It is known from branches further up the tree and archaeological results that haplogroup R arose in Asia. We can presume that U106 was founded somewhere late in this migration from Asia to Europe.

(2) Kurgan hypothesis

Key mutations often arise during population expansion events. These will typically shortly precede (or occur during) migration events when one group takes over another. The important Asia–Europe migration taking place around the time of U106's formation was the Kurgan expansion out of the Russian Steppe.

(3) Upstream SNPs

The geographical median of SNPs between M269 and U106 follows an east–west trend. We can use this to infer that the M269→U106 sequence follows a migration from the east to the west. The Kurgan expansion is the only known, major migration that fits both the likely range of dates and the east–west movement. In addition it appears to arise from the trans-Ural area near concentrations of groups further up the haplogroup R1b tree (e.g. V88).

(4) Urheimat

The Gimbutas interpretation of the Kurgan hypothesis credits the Kurgans with the introduction of the Indo-European language family to Europe. The origin of this language is referred to as the *Urheimat*, and is generally considered to have been in the period 4200–3500 BC. Its location is unknown and may be anywhere from the Causcasus to the trans-Ural region shown here.

(5) M269

M269 formed around 4100 BC, but the uncertainty in its age is roughly 600 years, so identifications with a particular culture are highly speculative. Kurgan wave 1, migration from the Volga to the Dneiper, took place around 4500–4000 BC and could be an early origin for M269. Wave 2 probably occurred from the Maykop culture (3700–3000 BC, indicated in cyan). The high variance and unusual M269 population found in this area (Hovhannisyan et al. 2014) leads us to prefer this for an origin for M269.

(6) M269→L23→L51

Hovhannisyan et al. (2014, fig. 6) notes a dissimilarity between the Ossetian, Azerbaijani, Turko–Armenian and European M269 populations. L51 does not clearly appear in the ancient DNA results itself. However, ancient DNA from the Yamnaya culture (modern day Russia and Ukraine) show that L23 and its subclade Z2013 dominate here. Additionally, FTDNA shows M269+ L23- and L23+ L51- tests are much more focussed on the Black Sea than L51+ tests. We interpret this as indicating L51 formed just after the migration started heading from the Black Sea towards western Europe.

(7) M269→L23→Z2013

L51 and its brother clade, Z2013, seem to have entered eastern Europe along with L51, but Z2013 does not seem to have participated much in the subsequent expansion west. PF7580 and its subclades are clearly concentrated in the Turkish highlands; CTS7822, particularly CTS9219, is closely associated with the Balkan peninsula; L277 is spread around eastern Europe; while CTS7763 may be Balkan or Anatolian. Meanwhile, ancient DNA from Haak et al. (2015) shows a very strong expansion of Z2013 throughout modern-day Russia. These data suggest Z2013 went north, south and east, while L51 went west.

(8) Caucasian populations

There are few cases where specific SNPs have been tested among people from the Caucasus, at least at FTDNA. Two M269 testers are L23→Z2013 and L23→L51→P310. Combined with the Hovhannisyan and ancient DNA (Haak et al.) results, we suggest they are probably mostly L23→Z2013.

(9) M269→L23→Z2013→PF7580 and the Hittites

The variance of PF7580 in Turkey and Syria suggests a coalescence age of 3000–5000 years. The Hittites are thought to have arrived in Anatolia before 2000 BC. On this basis, we ascribe the origin of PF7580 to a Hittite population.

(10) L51→PF7589: migration into Europe

L51 through to P311 represents an ~800-year gap in our knowledge. This may represent a hiatus in the east–west population movement which could be traced by historical cultures. This lack of structure makes it difficult to tell what went on in this ~800-year period.

Some information can be found from related clades. L51→PF7589 shares the European focus of L51→P311, but is not widely found in Germany. The median location of PF7589 in FTDNA tests is close to Salzburg, but the east–west migration means the split between P311 and PF7589 is likely to be further east. So although L51 may represent the population that launched into Europe, they perhaps (initially) did not get very far.

(11) The path into Europe: North or South?

The path our ancestors took into Europe isn't well defined. There is a general preference for a path up the Danube river, which leads directly from the Black Sea to Germany, where P311 is first found in the archaeological DNA, and where there is the easternmost substantial(!) concentration of P311 in the present population. This also fits better with the distribution of downstream clades.

Conversely, there is substantial reason to believe the path was actually over the northern side of the Carpathian mountains, through Poland. This is the area through which the Corded Ware culture spread. Recent results from ancient DNA have provided more R1a results in earlier times, suggesting R1b did not participate widely in the Corded Ware culture to begin with. However, this is based on a very few ancient samples and may not stand up to long-term scrutiny.

A northern route would be slightly preferred by the ancient DNA results, which indicate significant R1a concentrations in southern Germany, and R1b in northern Germany and Poland. This also sits better with the only ancient U106 result so far in southern Sweden.

(12) Early P311 and arrival in Germany

P311 splits into U106 and P312. This split probably occurred sometime during the march westwards. If our ancestors took a northern route, the lack of P311 in Poland (where R1a dominates) suggests it cannot have been further east than the Germano-Polish border. If our ancestors took a southern route, the easternmost likely place is Austria. The founding of P311 itself may have been slightly earlier, and considerably further east.

Either way, P311 seems to have been present in Germany around 2700 BC, around the time that the Corded Ware culture, and more specifically the Single Grave Culture, were setting up shop there… perhaps quite literally, given the Baltic amber trade. The Single Grave Culture is one point considered for the start of the Bell Beaker culture. Ancient DNA shows P311 played a significant part in the Bell Beaker culture in Germany, and the spread of the Bell Beakers may have been instrumental in spreading P311 throughout western Europe.

(13) P312 and U106

P312 is U106's bigger (though not necessarily older) brother. It's worth considering how P312 fared after the P312–U106 split. Early P312 is found at the Quedlinburg site in central Germany, and remains from its subclade, U152, have been found in south-eastern Germany. Both these individuals were buried in Bell Beaker culture fashion.

U106 now appears more commonly in Germany and Scandinavia. By contrast, P312 dominates in most of western Europe. U152 is found mostly in Switzerland and Italy, DF27 in Iberia, and L21 among Brythonic peoples. P312 therefore appears to have travelled south and west, while U106 mostly either stayed in Germany or moved north.

(14) The foundation of U106

U106's earliest origins can probably be traced to Germany, although Austria is also a possibility for the southern route. We have placed this foundation at around 2650 BC, give or take a few centuries, which allows U106 to form part of both the Single Grave and Bell Beaker cultures, plus the pre- and proto-Cetlic cultures that followed them.

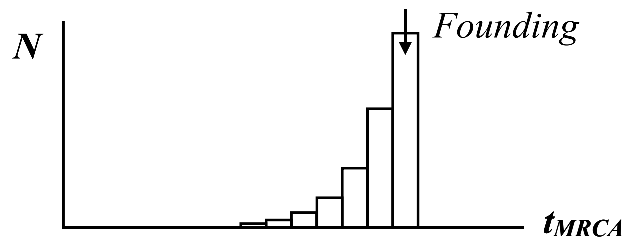
The U106 ancient DNA from Lille Beddinge in Sweden (skeleton RISE98) show that our ancestors weren't idle, and kept moving. While the particular line of RISE98 seems to have died out, we can presume that U106's ancestors formed part of the Nordic Bronze Age too.

Migrations from STRs

MIGRATION PATHS FROM STRs

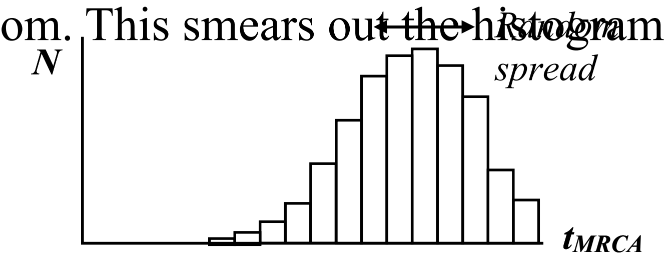
Migrations in recent times can be traced through histograms of times since the most-recent common ancestor (TMRCA) from STRs. This is an extension of the previous STR-dating method that can be used to disentangle geographies and migrations.

The principle works by measuring the TMRCA for every pair of men within a clade and comparing the distribution of people. In an iideal but growing population, where a man sires two sons, who each have two sons, who each have two sons, etc., the histogram will look like this:

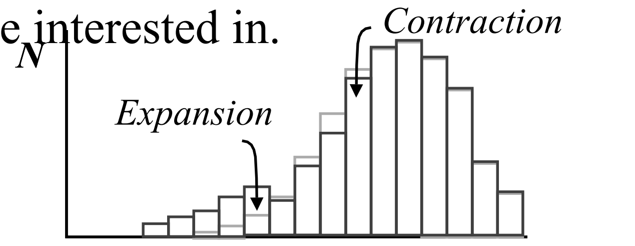


All men are related to the founding father. Half of men are related through one son, half through the other, so half of the table of TMRCAs will be a generation younger. Of each of those halves, half will be related another generation down, etc. So we end up with a histogram that halves with each generation, like the one above.

Generations aren't exactly the same length and the mutation process is random. This smears out the histogram:



In the real world, the expansion and contraction of populations occurs in response to external and internal events. This means that clumps form in the histogram during periods of population expansion, and gaps appear during population contraction. This modifies slightly the shape of the histogram. It is these bumps and voids that we are interested in.



These effects are subtle, and best illustrated through a real-world example. Here, we consider two examples:

U106>Z381>Z156>DF98 ,

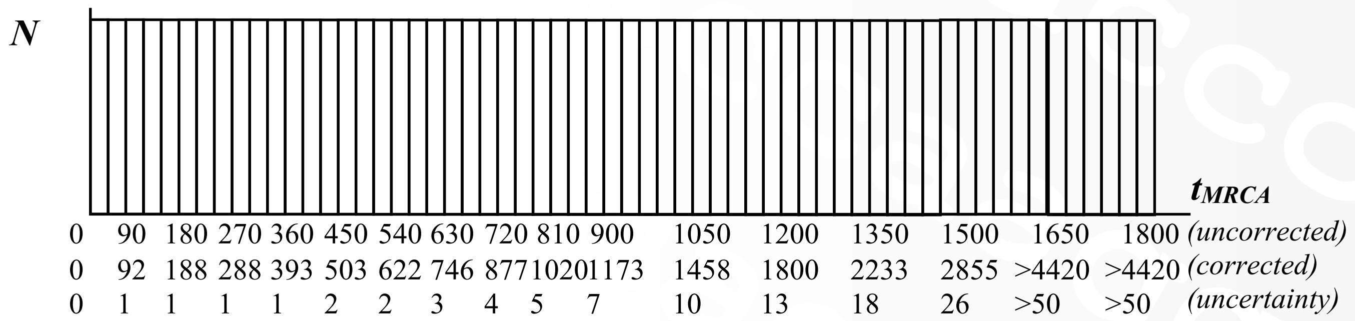
and

U106>Z381>Z301>L48>Z9>Z30>Z2>Z7>Z8 .

These have very different backgrounds. DF98 concentrates in the Rhine valley and is known to have at least two Norman or Norman-era migrations. Z8 concentrates in the Low Countries, Germany and England and looks much more Germanic. We expect to see differences in their structure. First, however, we have to perform a calibration.

HISTOGRAM TMRCA CALIBRATION

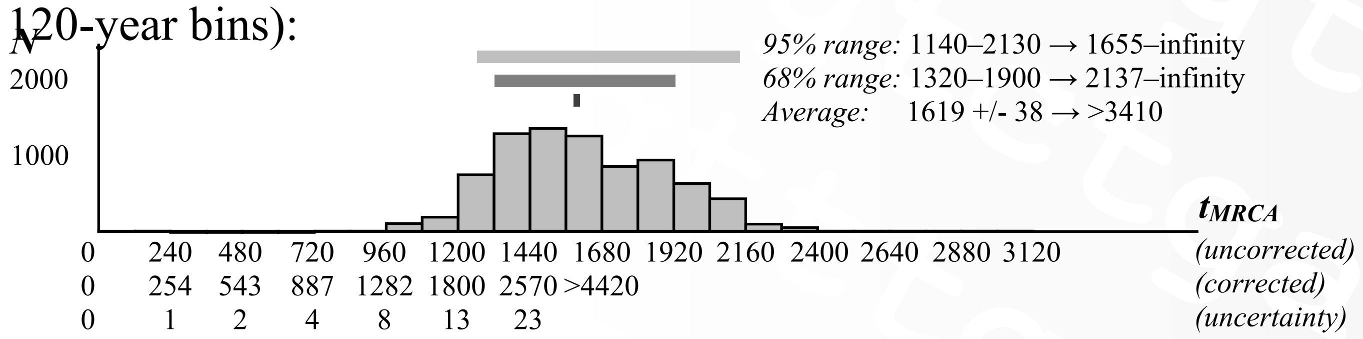
Earlier, we discussed how the STR data were calibrated against the SNP data for the infinite alleles method. Corrections to the infinite alleles method become significant after a few centuries, and the method becomes largely useless much after 2000 years ago. This means we expect the following translation of STR-based ages to reality:



Note how the correspondence is lost after ~3000 years as the ages tend to infinity as our correction function fails to fit. The uncertainty in this case is in the accuracy of our fitting function, and doesn't take random spread into account.

CHARACTERISING RANDOM SPREAD THROUGH INTER-CLADE TMRCA DISTRIBUTIONS

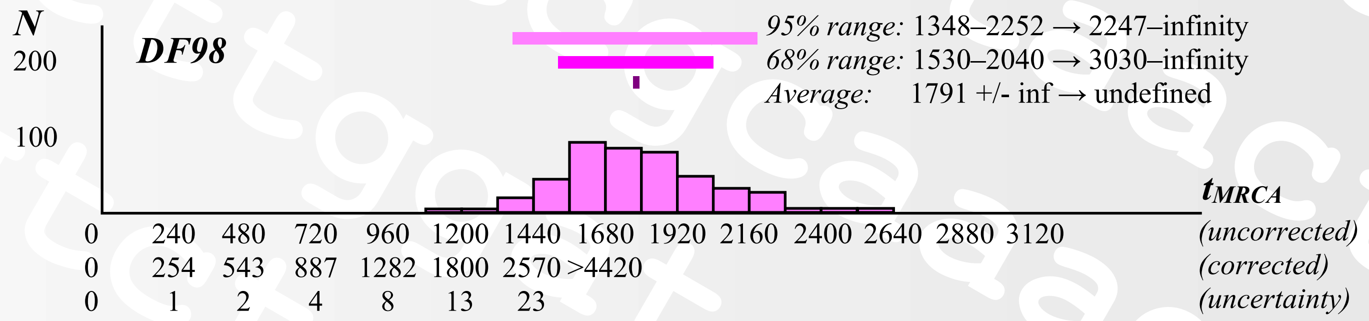
We can now take our example clades, DF98 and Z8, and measure their characteristic intrinsic spread. This spread is a function of testing depth (number of mutations tested) and age of population (number of mutations accumulated). DF98 and Z8 are last related by Z381, around 4400 years ago. Comparing the STR TMRCAs for DF98–Z8 pairs*, we arrive at the following histogram (binned into 120-year bins):



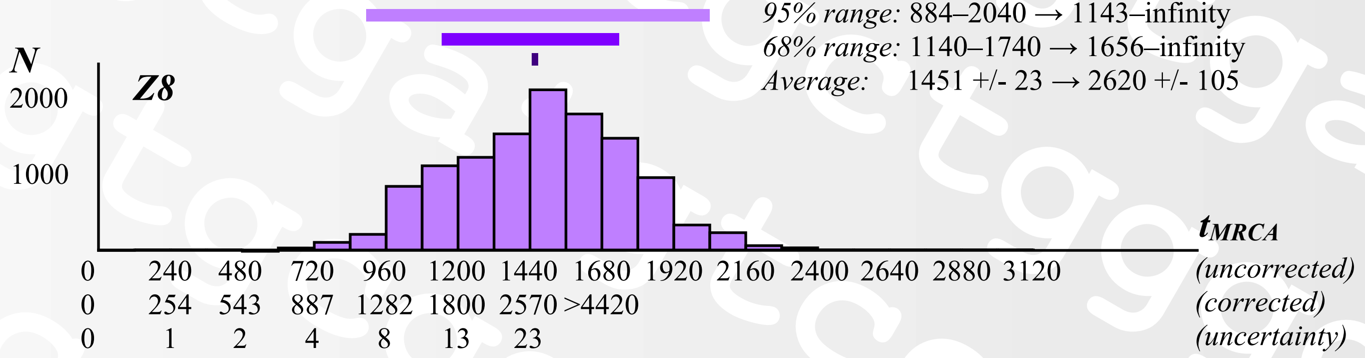
This histogram exemplifies the limitations of this approach. The true relationship age is around 4400 years ago for all pairs in this histogram. Three randomly sampled pairs from this histogram are most likely to have relations predicted to be close to 1320, 1619 and 1900 years ago (a typical uncertainty of 290 years). These ages would be corrected to circa 2137, >3410 and >4420 years. DF98 is predicted to be ~3600 years old, and Z8 to be ~2400 years old. Any migrations between the ~4400-year-old Z381 foundation and the ~3600-year-old DF98 foundation will be lost in this random spread. Migrations around the Z8 foundation might be recoverable, but only if they are very significant. The limitations of this method probably lie around 1000-2000 years ago, depending on the number of testers and scale of the migration.

(NB: Only Z8>Z334 tests were used in this analysis due to the large number of Z8 testers and the computational power required, which scales as N_{testers}^2).

The same analysis can be performed on DF98 and Z8 themselves, using their subclades, S1911 and S18823, and Z1 and Z11, respectively.



Despite the correction, the age of DF98 is not predicted: it is older than age the infinite alleles method is stable over. The spread of the histogram, however, has reduced from 290 years (uncorrected) for Z381 to 255 years (uncorrected) for DF98. This suggests that, at best, we can expect an accuracy of ~200 years in the dating of migrations within DF98. This is roughly what we would expect, as it is similar to the STR mutation rate (~1 per 140 years at 67 markers). Structure in the histogram younger than ~2000 years ago can probably be dated with relative accuracy.



In the case of Z8, the age is slightly over-predicted at ~2620 years instead of ~2400 years, but agrees well once the uncertainties are considered. Despite having a younger age, the spread of TMRCAs remains at ~300 years. Due to this spread, we can't use this method to understand any structure in Z8 before about 1300 years ago.

In these *inter*-clade histographs of DF98 and Z8, there is a nearly Gaussian ("normal" or "bell-curve") distribution of values with a characteristic spread. However, the distribution isn't quite Gaussian: e.g. Z8 has a 'bulge' around 1500 (corrected) years ago. These imperfections can reflect parallel or opposing mutations, typically from early in the history of that clade. In this case, it is due to a comparative lack of mutations in the Z1>Z344>14436052 and Z338>Z11>Z8175>…>FGC12059 clades. This will lead to artefacts in the *intra*-clade spreads that we will use to work out relationships.

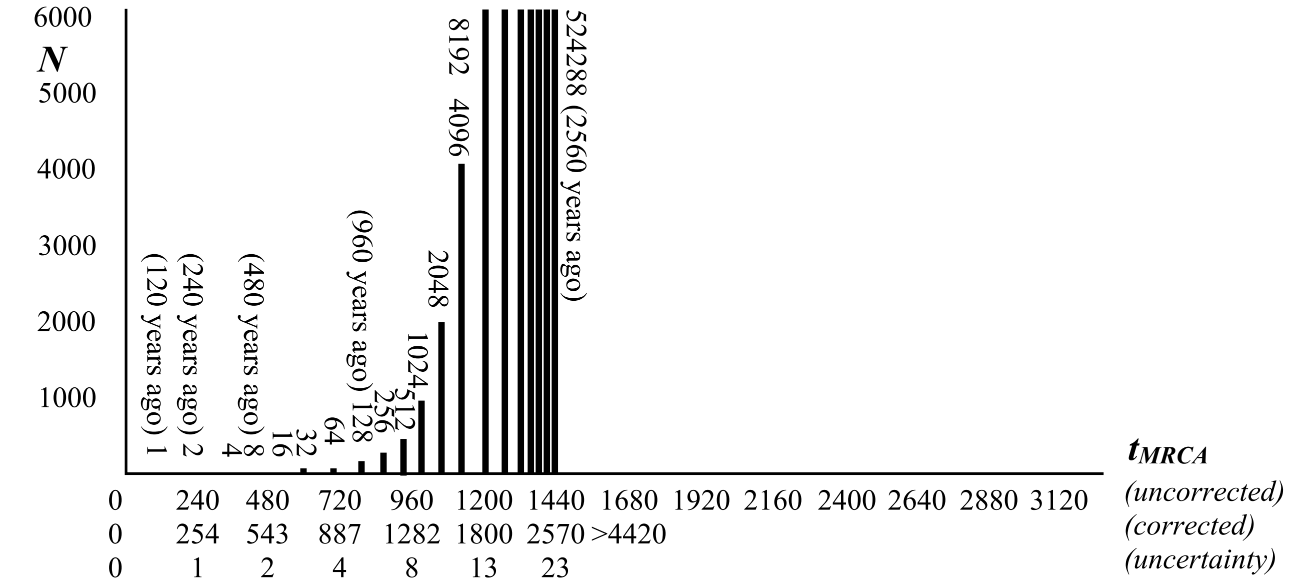
Experiments on several clades has shown the random spread can be characterised fairly consistently as ~200 years for young clades, rising to ~300 years for older clades. but varying between the two values, depending on the clade in question. We can therefore adopt a 200~300 year time period as the typical random spread in an uncalibrated TMRCA distribution. Other features are therefore likely to be due to internal structure.

The examples above show that we are limited to tracing migrations younger than 1000~2000 years, depending on the clade involved. Migrations older than this will be lost in the noise of the clade, but may still be recoverable using other methods like geographical distribution.

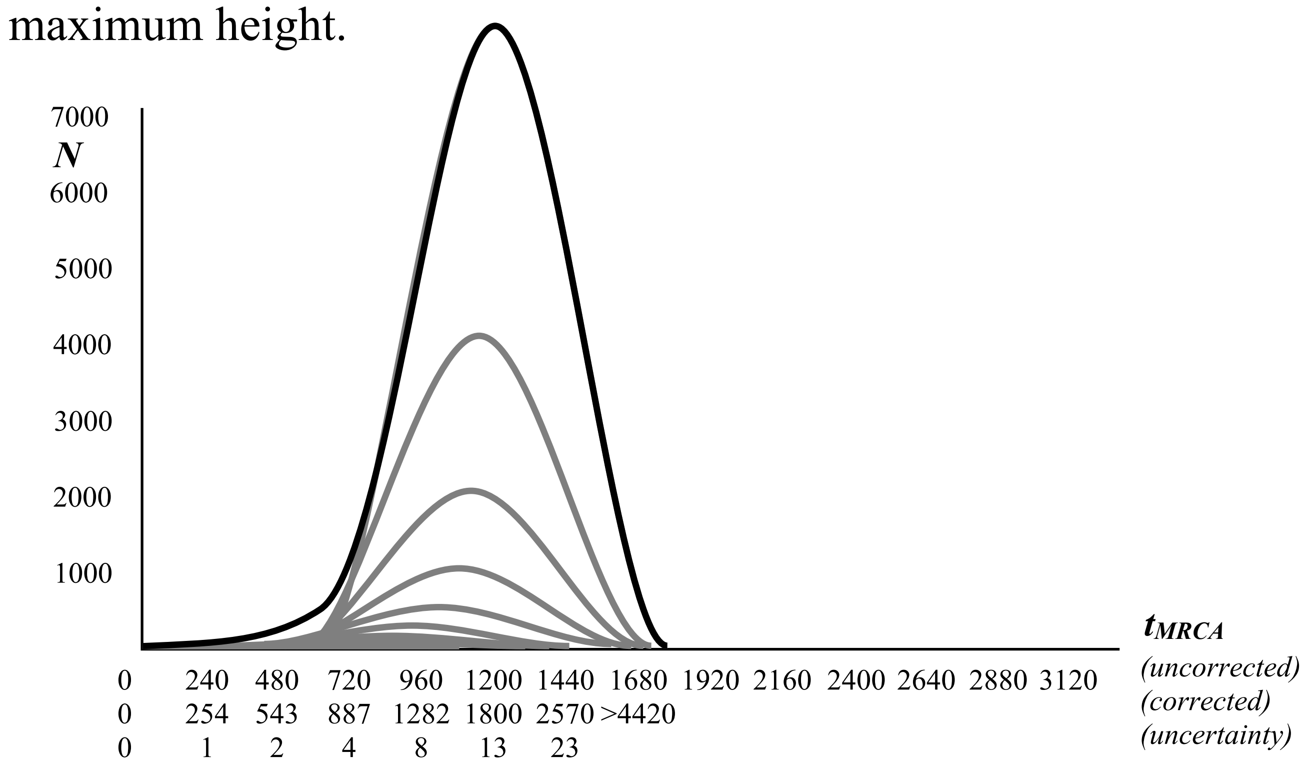
INTRA-CLADE HISTOGRAMS: EXPECTATIONS

Turning our attention to the *intra*-clade TMRCA histograms, we can form an expectation of what one should look like.

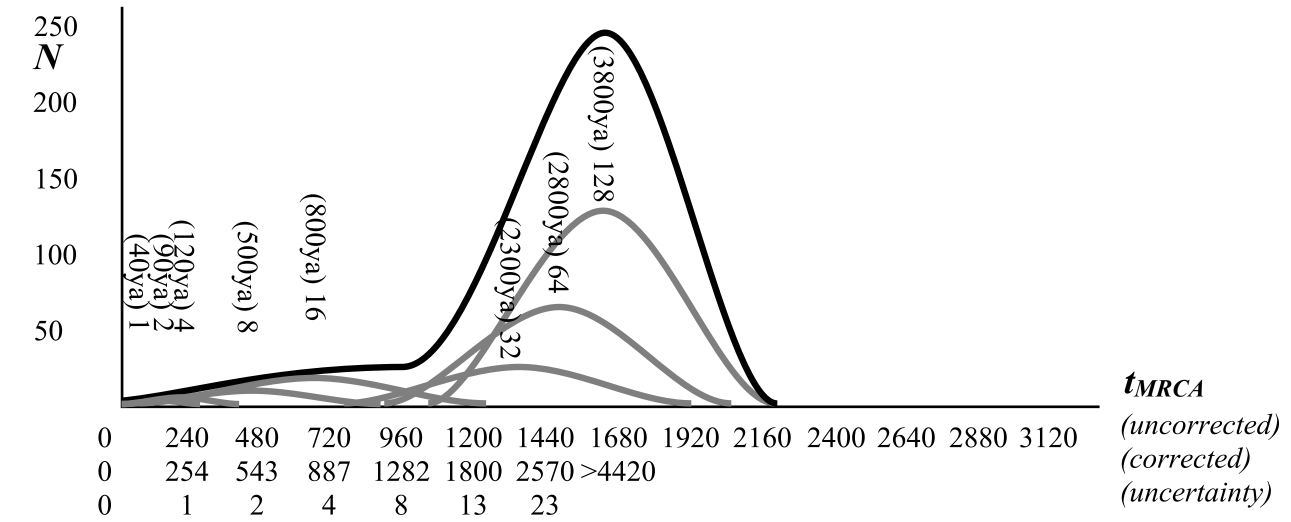
In the simplified population presented previously, we have a slowly growing population, with a new branch forming every 90-140 years or so, depending on where 67 or 111 markers are being tested (let's say 120 years). Half of the population goes into each branch, so half of the MRCAs are 140 years closer to the present than the other half. Of that closer half, half are another 140 years closer to the present, etc., so the histogram of TMRCAs halves every 140 years of real (corrected) time, following a 1/t² law.



These will each be smeared out with a Gaussian of 200-300 years in width in *un*corrected time (width being the width at half the maximum height.



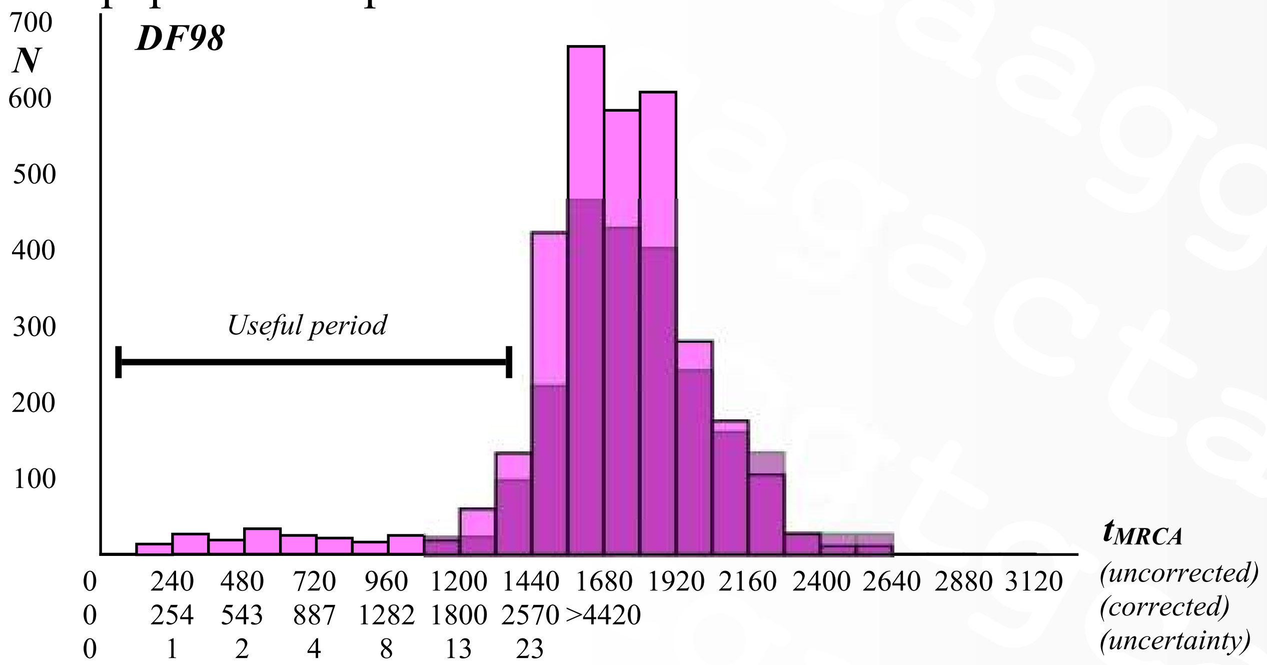
But in the real world, populations grow and shrink, so we can instead expect to find a new branch forming every time that population doubles.

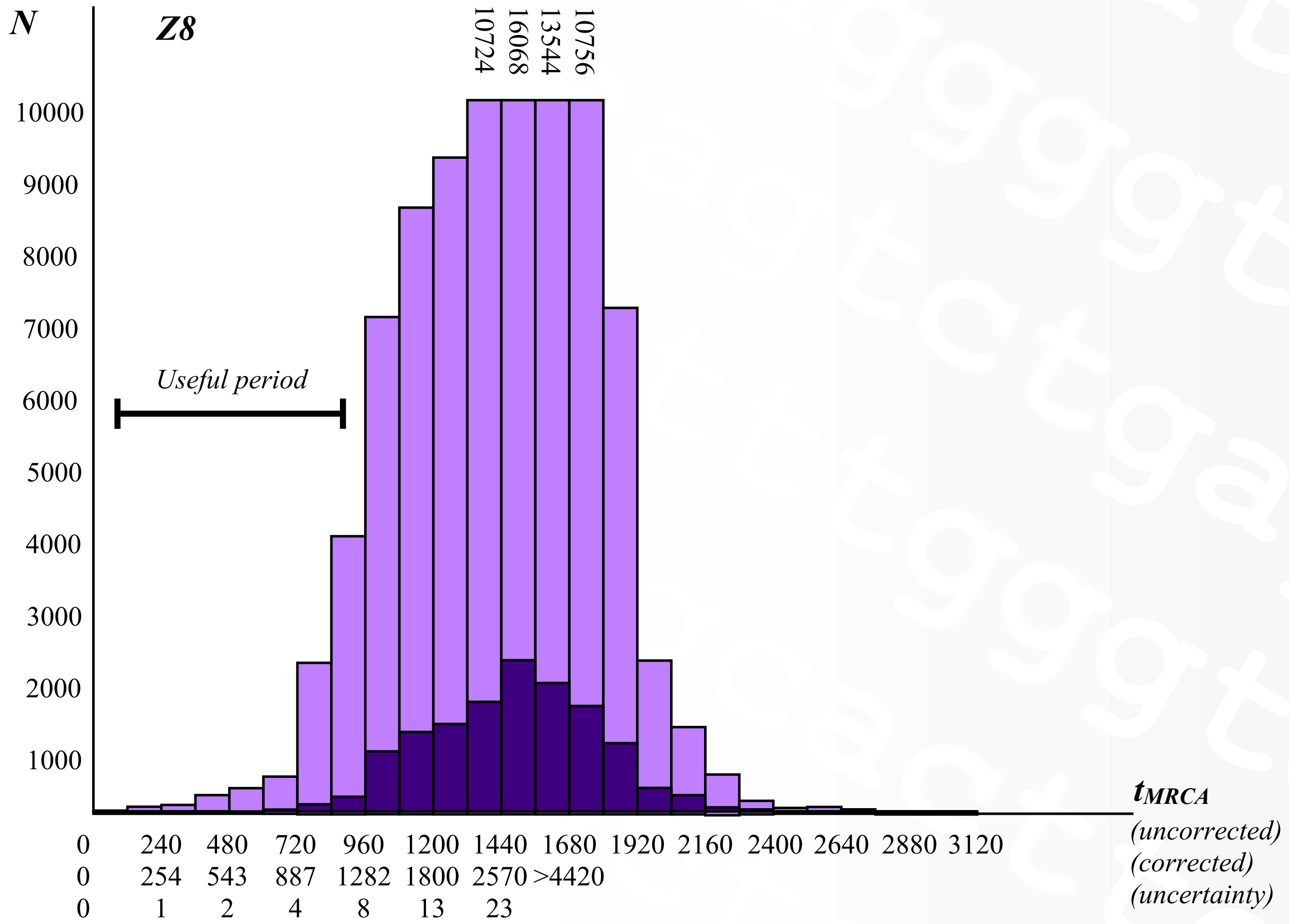


INTRA-CLADE EXAMPLES

In reality, things will be rather more complicated, as large branches are resiliant to population shrinkage, while small branches aren't, and branches occur in a random process, not a strictly timed fashion. But the basic structure is of a large peak when the population first formed, followed by a tail containing useful information about when that population grew in size.

We can now look at the real-world examples for Z8 and DF98 to see what they reveal. This should show us roughly what happened with population expansion and contractions.



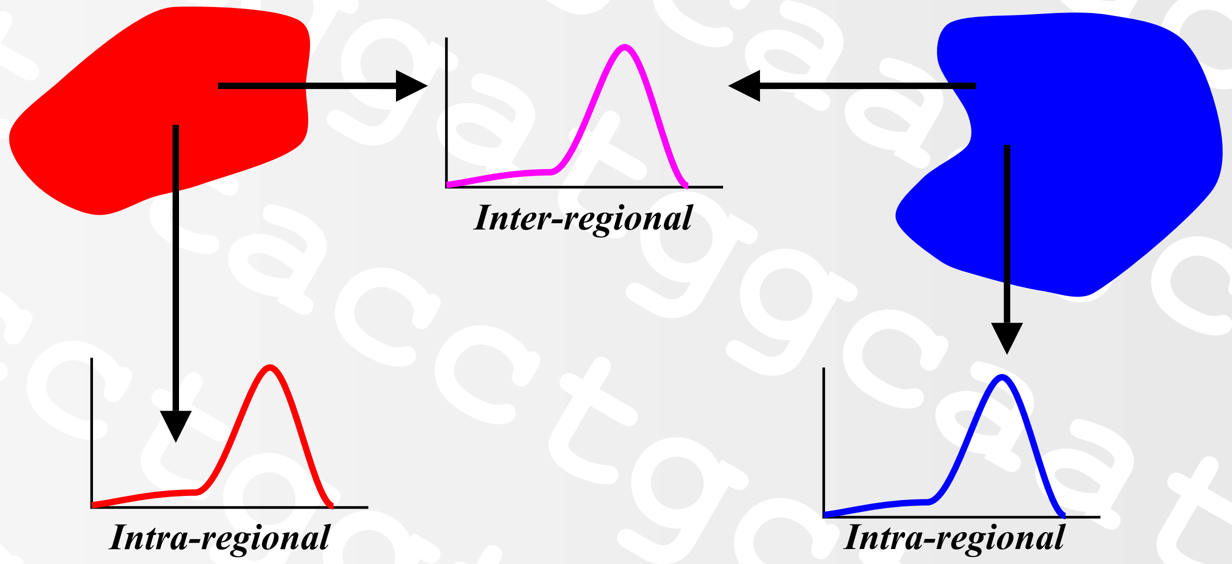


Both graphs show a strong peak, then a weak tail to modern times. DF98 has a long and markedly constant tail, showing continued and even accelerating population expansion for the last 1800 years. By contrast, the rapidly falling tail of Z8 shows that the modern growth of Z8 is at most at the background rate, and significant growth of Z8 above the general population probably ceased at least 900 years ago.

Separating this useful "tail" from the bulk of the TMRCAs is not easy, so some care must be taken about assuming hard evidence from such distributions.

COMPARING GEOGRAPHICAL REGIONS

The major advantage of this test comes when we start to split up these intra-clade TMRCAs by region to provide *inter-regional* and *intra-regional* TMRCA histograms for different regions or countries.



Peaks in an *intra-regional* TMRCA histogram indicate a growth of that clade in that country: e.g. a growth in England around 900 years ago could indicate the rapid growth from a Norman population. By contrast, peaks in an *inter-regional* TMRCA histogram indicate a migration: e.g. if there is a peak in the Anglo-French inter-regional histogram around 900 years ago, we can attest this to a migration between England and France (or vice versa) around that time.

Used together, these techniques can be used to trace migrations. For example, if no peak (or peak much older than 900 years) is seen in the French histogram, we can infer the direction of migration was from France to England. If a younger peak exists in the French histogram, we can infer migration from England to France.

The imprint of peaks in the population of origin should show up in the destination region too. For example, we could expect a peak in the inter-regional TMRCAs of Scotland and of Ireland around 900 years ago, as although the Norman influences from England didn't arrive there straight away, they still brought their Norman signatures with them when they came.

As seen in the previous example, these peaks are very hard to detect and can be very ambiguous. Treated with caution, and with careful examination of the underlying and associated evidence, they can prove useful in tracing past migrations.

VARIANCE AS AN ORIGIN INDICATOR

A final piece of evidence we can use is the position of the peak. The oldest population should have the oldest average TMRCA. Often this effect is hard to identify, but can be very useful where the "founder effect" exists: a slow-moving migration where one or a small number of people from a clade move into an area long after that clade has been founded. In these cases, the average TMRCA for that region may be considerable younger than the original population. The use of genetic variance as an indicator of origin is a well-known tool. However, stastistical spread can cause substantial differences on its own, so care must again be taken when using this method.

Factsheet: U106

Dr. Iain McDonald
on behalf of the U106/S21 group

U106

Number of testers at 67/111 markers (N) = 997/475
Variance at 67/111 markers (σ) = 20.10/32.56
Mean TMRCA from infinite alleles (μ) = 1917 years (uncorrected)
Median TMRCA from infinite alleles (M) = 1882 years (uncorrected)
Standard deviation among TMRCAs (s) = 363 years (uncorrected)
The same symbols are used in the regional comparisons below

Likely origin: Germany, 2600 BC
Possible earliest association(s):
Corded Ware Culture, Single Grave Culture
Protruding-Foot Beaker Culture, Bell Beaker Culture
Primary regions:
Germany, Low Countries, north/east France, south/east England

Migrations:

Proof-of-concept version. Currently missing Wales, Switzerland, Croatia and Cyprus

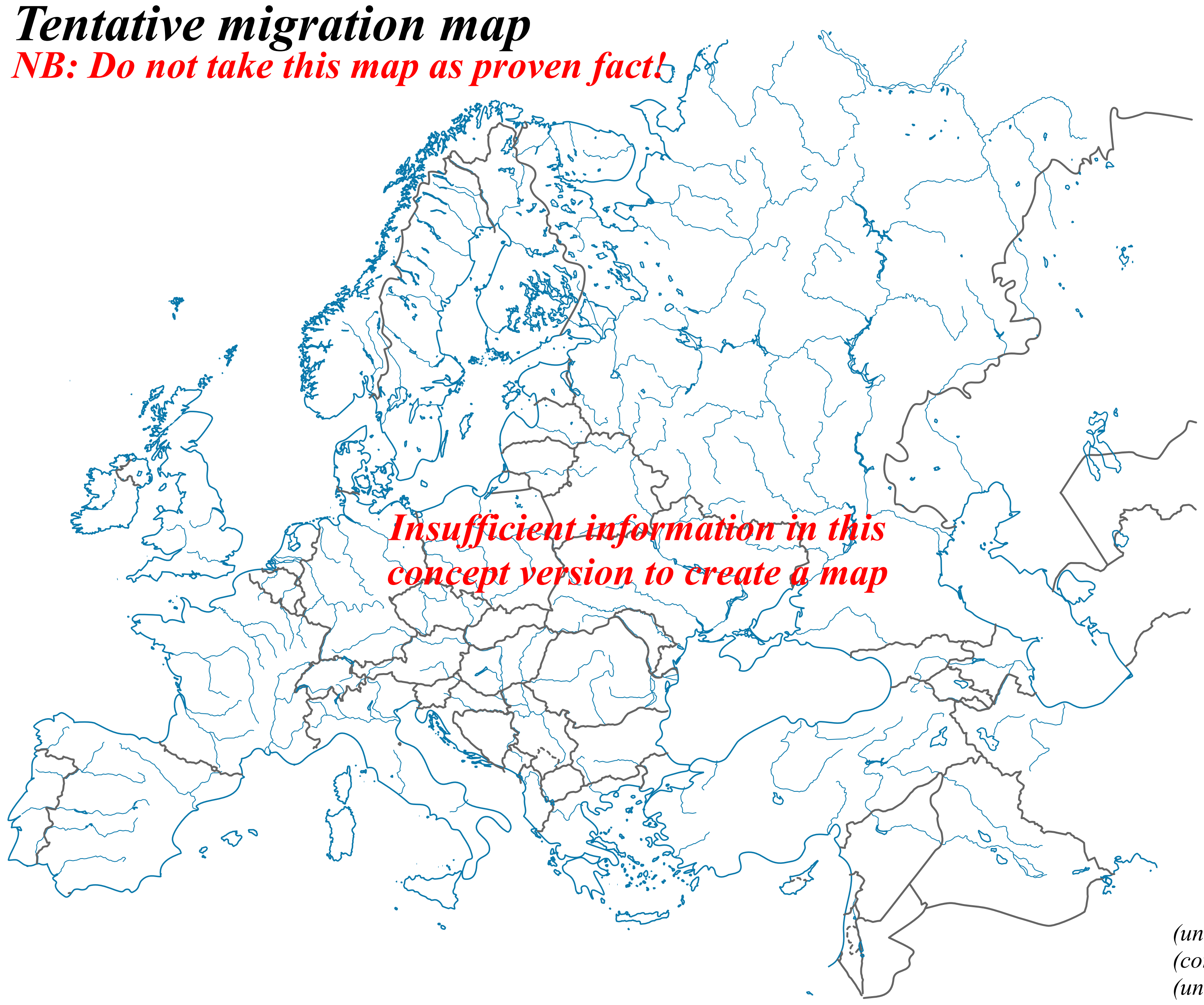
What to look for in histograms: For each regional pair, a histogram like that on the right is shown. Its height is generally in proportion to the number of pairs of people in each group (*A* x *B*). The histogram will generally tail off from the main distribution as the one on the right does. Bumps on top of this tail indicate migration between the two countries occurring shortly after this point.

The histogram displays dates between 0 and 1800 calibrated years ago. Three regions are indicated: post-colonial (c. 1500 AD to present), post-Norman (c. 1000 AD to 1500 AD) and pre-Norman c. 200 AD to 1000 AD).

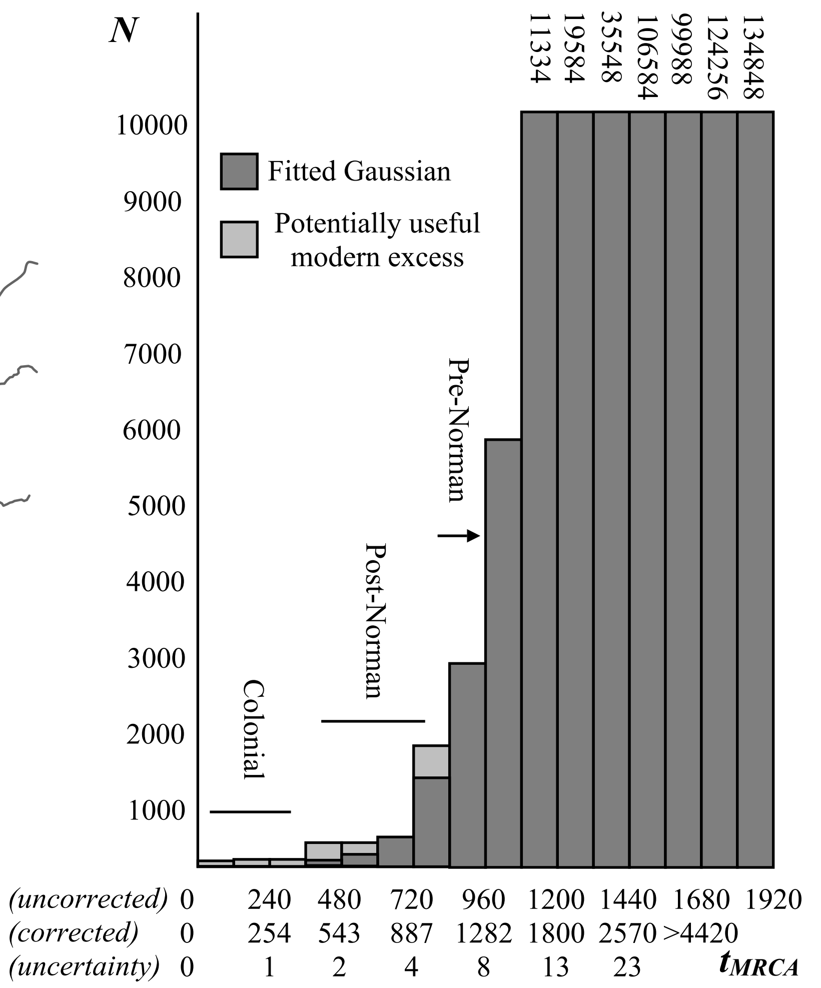
Ireland *N*: 113/59 σ. 20.64/33.38												
σ. 40.92/64.59 *μ=1914; M=1884* *s=359*	**Scotland** *N*: 98/56 σ. 19.79/30.03											
σ. 39.49/64.84 *μ=1895; M=1875* *s=354*	σ. 38.75/61.28 *μ=1903; M=1875* *s=355*	**England** *N*: 347/173 σ. 18.47/30.20										
σ. 42.54/67.93 *μ=1920; M=1882* *s=381*	σ. 42.04/64.71 *μ=1938; M=1887* *s=382*	σ. 40.20/64.10 *μ=1910; M=1875* *s=392*	**Low Countries** *N*: 55/17 σ. 23.89/34.35									
σ. 42.84/72.04 *μ=1951; M=1893* *s=326*	σ. 42.53/69.38 *μ=1973; M=1945* *s=312*	σ. 40.62/68.49 *μ=1936; M=1887* *s=330*	σ. 43.54/71.57 *μ=1957; M=1892* *s=358*	**France** *N*: 41/23 σ. 20.82/35.91								
σ. 44.27/69.99 *μ=1906; M=1882* *s=346*	σ. 44.24/66.27 *μ=1917; M=1884* *s=341*	σ. 42.33/65.68 *μ=1893; M=1871* *s=355*	σ. 48.11/70.92 *μ=1926; M=1885* *s=384*	σ. 44.79/72.38 *μ=1901; M=1874* *s=330*	**Iberia** *N*: 10/4 σ. 21.71/29.52							
σ. 41.44/70.22 *μ=1923; M=1883* *s=347*	σ. 40.87/66.89 *μ=1928; M=1888* *s=356*	σ. 39.04/65.53 *μ=1900; M=1876* *s=345*	σ. 44.30/68.81 *μ=1926; M=1880* *s=379*	σ. 42.29/71.82 *μ=1957; M=1893* *s=313*	σ. 43.74/67.99 *μ=1946; M=1900* *s=327*	**Italy** *N*: 15/3 σ. 18.37/26.60						
σ. 39.11/65.74 *μ=1968; M=1926* *s=352*	σ. 38.78/62.59 *μ=1968; M=1936* *s=337*	σ. 37.18/62.44 *μ=1945; M=1910* *s=361*	σ. 42.75/66.73 *μ=1994; M=1989* *s=348*	σ. 39.53/68.83 *μ=1993; M=1987* *s=317*	σ. 41.00/62.90 *μ=1947; M=1943* *s=375*	σ. 39.14/65.04 *μ=1985; M=1988* *s=346*	**Eastern Europe** *N*: 24/13 σ. 17.86/29.61					
σ. 42.03/70.14 *μ=1928; M=1879* *s=353*	σ. 41.62/65.42 *μ=1954; M=1887* *s=356*	σ. 39.72/64.86 *μ=1911; M=1876* *s=356*	σ. 44.94/68.54 *μ=1933; M=1884* *s=385*	σ. 42.20/72.08 *μ=1944; M=1893* *s=335*	σ. 45.03/68.08 *μ=1912; M=1885* *s=372*	σ. 41.12/67.38 *μ=1933; M=1882* *s=379*	σ. 39.13/63.27 *μ=1928; M=1894* *s=375*	**Cz/Sl/Au/Hu** *N*: 20/5 σ. 19.81/30.29				
σ. 40.26/67.14 *μ=1912; M=1881* *s=353*	σ. 39.64/64.03 *μ=1926; M=1887* *s=348*	σ. 37.93/63.66 *μ=1890; M=1872* *s=358*	σ. 43.72/68.32 *μ=1910; M=1875* *s=394*	σ. 40.97/70.82 *μ=1934; M=1888* *s=326*	σ. 42.71/68.35 *μ=1872; M=1870* *s=338*	σ. 39.58/68.00 *μ=1921; M=1884* *s=376*	σ. 37.47/64.79 *μ=1947; M=1907* *s=347*	σ. 40.47/68.14 *μ=1912; M=1882* *s=365*	**Germany** *N*: 110/51 σ. 19.13/32.65			
σ. 40.21/61.54 *μ=1993; M=1986* *s=352*	σ. 39.22/57.83 *μ=1975; M=1925* *s=349*	σ. 37.55/58.59 *μ=1962; M=1896* *s=350*	σ. 43.42/63.11 *μ=2000; M=1982* *s=377*	σ. 41.29/65.77 *μ=1997; M=1984* *s=333*	σ. 42.62/66.04 *μ=1976; M=1891* *s=357*	σ. 39.29/63.29 *μ=1998; M=1997* *s=348*	σ. 37.27/61.10 *μ=2034; M=2007* *s=366*	σ. 39.90/63.50 *μ=2025; M=1990* *s=352*	σ. 38.50/61.20 *μ=1970; M=1927* *s=358*	**Denmark** *N*: 17/3 σ. 18.01/22.68		
σ. 43.26/71.38 *μ=1948; M=1889* *s=351*	σ. 42.21/67.38 *μ=1960; M=1911* *s=359*	σ. 41.13/67.81 *μ=1935; M=1885* *s=363*	σ. 46.69/72.59 *μ=1940; M=1886* *s=388*	σ. 43.89/75.13 *μ=1988; M=1985* *s=336*	σ. 45.67/73.91 *μ=1935; M=1885* *s=372*	σ. 42.25/73.45 *μ=1945; M=1901* *s=353*	σ. 39.72/69.64 *μ=1989; M=1943* *s=343*	σ. 43.65/73.42 *μ=1948; M=1894* *s=376*	σ. 41.42/69.99 *μ=1928; M=1885* *s=359*	σ. 41.05/63.00 *μ=2016; M=1993* *s=356*	**Fenno-Scandia** *N*: 82/30 σ. 20.64/34.01	
σ. 42.55/68.32 *μ=1968; M=1925* *s=344*	σ. 42.43/65.29 *μ=1982; M=1960* *s=353*	σ. 40.86/65.41 *μ=1964; M=1925* *s=363*	σ. 46.40/69.65 *μ=1983; M=1948* *s=383*	σ. 43.62/72.13 *μ=2015; M=1994* *s=322*	σ. 44.58/66.43 *μ=1977; M=1987* *s=410*	σ. 43.01/68.84 *μ=2003; M=1993* *s=349*	σ. 39.11/64.86 *μ=2019; M=2007* *s=401*	σ. 42.47/66.92 *μ=1970; M=1928* *s=393*	σ. 41.07/67.44 *μ=1974; M=1935* *s=351*	σ. 41.28/63.45 *μ=2083; M=2021* *s=371*	σ. 43.49/71.74 *μ=1991; M=1983* *s=362*	**Poland&Baltic** *N*: 43/18 σ. 20.42/32.34

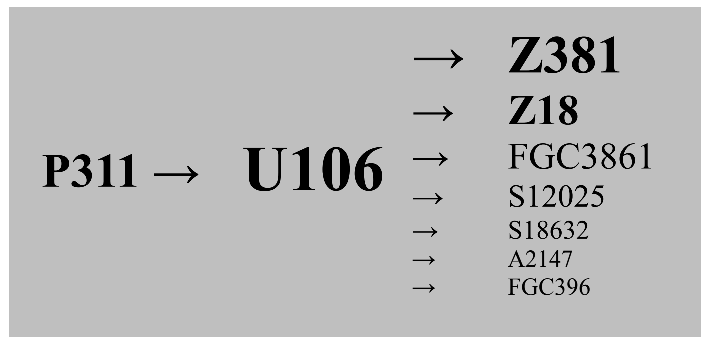
Tentative migration map

NB: Do not take this map as proven fact!


Insufficient information in this concept version to create a map

Infinite alleles TMRCAs





Notable histogram findings:

Scotland–Ireland: migrations between the two seem common shortly after 550 years ago. This is probably linked to the Plantations shortly after 1600 AD.

England→Scotland/Ireland: possibly significant migrations occurring in the post-Norman era. Hard to separate.

Low Countries→British Isles: significant relations in immediate pre-Norman era. Could be linked to many migrations between the Saxon, Viking and Norman eras.

France→British Isles: significant excess around 1000–1300 years ago, likely linked to the Norman migrations.

Germany/Denmark→British Isles: significant excess shortly before 1300 years ago, likely linked to the "Anglo–Saxon" (post-Roman) migrations.

Scand.→Low Countries & Germany: possible Viking-era excess? Not readily apparent in the British Isles.

Poland/Baltic: strong suggestions of a migration around 900 years ago. Source is difficult to identify, but suspect Viking influence (note: not just the Z159 cluster).