U106 explored: its relationships, geography and history
The 2015 report to the U106 group (September update)
Principal investigator: Iain McDonald

Contents

Background to genetic testing and history before U106 2
Foreword & basics of Y-DNA testing 2
Characterisation and quality control of next-generation tests 3
Deep ancestry of U106: 200,000 BC to 2500 BC 4
The route out of Africa to M269 (200,000 BC to 4000 BC) 5

Age estimation and structure within U106 6
Methods of age estimation from next-generation tests 6
Cross-calibration of ages from STR and SNP tests 9
Structures within the U106 tree 11

Present-day geographical makeup of U106 12
Present-day distribution of U106 in Europe 12
Present-day distribution of U106 in the British Isles 13

Spread and ancestry of U106 14
Methods for determining the source, direction and date of migrations 14
Evidence from archaeological DNA 15
Hypothesised migration of U106 ancestors into Europe 19
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The dates I compute here were outdated before I finished compiling the information.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

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. 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 (TM RCA). 

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

could which mutate to:

DYS1234 = 5 TACATACATACATACATACATACATACATACATACATACATAC

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 tests, we can build a relationship tree for families, e.g.:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYS389</td>
<td>DYS390</td>
<td>DYS388</td>
<td>DYS391</td>
<td>DYS392</td>
</tr>
<tr>
<td>389</td>
<td>390</td>
<td>391</td>
<td>395</td>
<td>386</td>
</tr>
<tr>
<td>A:</td>
<td>13</td>
<td>24</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>B:</td>
<td>13</td>
<td>24</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>C:</td>
<td>13</td>
<td>24</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>D:</td>
<td>13</td>
<td>24</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>E:</td>
<td>13</td>
<td>24</td>
<td>14</td>
<td>11</td>
</tr>
</tbody>
</table>

we presume the group CDE are more closely related because of the DYS39=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 ↔ ATGCTGATGATCGC

deleted:

ATGCTGATGATCGC ↔ ATGCTGATCGC

or mutated:

ATGCGATGATCGC ↔ ATGCTGATCGC

Sites of these latter mutations are known as a 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 encompasses 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” 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 data 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.: 67287193 67287069 67286939 67286776

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

- 67287193 means that a mutation from A to G has occurred at position 67287193.

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:

<table>
<thead>
<tr>
<th>chrY</th>
<th>67287193</th>
<th>G</th>
<th>356</th>
<th>REJECTED</th>
<th>GT 0</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>chrY</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
<th>22178569</th>
</tr>
</thead>
</table>

There are 220 (3.2%) false negatives on the lower end of coverage boundaries, and 5 (0.07%) 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 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 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%).

**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 22219409 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.

**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 phylogenetic tree of the human population represents our current understanding of the way the human family tree has divided along its mate lines. This is a rapidly-evolving topic, and 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 Y-chromosome family 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 consilient 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 descended from him via our 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 relatives hiding among remote Africa tribes. Haplogroup BT arose in Africa about 70,000 years ago, when 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 extend 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 15,800 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 lie 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, archaeological remains show that there was extremely few haplogroup R men in Europe before 2000 BC, when membership of these SNPs and populations, both SNPs originate several thousand years before these terms are relevant.

(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 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.

The 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.

Homo sapiens

The Y-chromosomal Adam 
around 200,000 BC, East Africa

Haplogroup BT (R-M42) around 70,000 BC, East Africa

Haplogroup F (M-F89) circa 48,000 BC, South Asia

Haplogroup K2 (R-M264) before 43,000 BC, Asia

Haplogroup R (R-M207) before 22,000 BC, Asia

Haplogroup R (R-M343) circa 19000 - 15500 BC, Asian Steppe

Haplogroup Rb (R-M343) circa 3500 - 3500 BC, Asian Steppe

Haplogroup Rb1a1 (R-M343) circa 3500 - 3500 BC, Asian Steppe

Haplogroup Rb1a1a1 (P311) 3400 BC - 2150 BC, Central and Eastern Europe

Haplogroup Rb1a2 (M343) circa 3300 - 3500 BC, Asian Steppe

Haplogroup Rb1a2a (M343) circa 3300 - 3500 BC, Asian Steppe

Haplogroup Rb1a2a1a (M343) circa 3300 - 3500 BC, Asian Steppe

Haplogroup U106 (Rb1a2a1a1a) 3050 BC - 2100 BC

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 (phylogenetic) tree splits from its foundation down to the U106 branch. Different current and prehistoric branches are shown in the chart, which can be interpreted carefully.

The date of this expansion can probably lie 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, archaeological remains show that there was extremely few haplogroup R men in Europe before 2000 BC, when membership of these SNPs and populations, both SNPs originate several thousand years before these terms are relevant.

Haplogroup Frequencies in Europe

The following data give the number and percentage of various levels between R-M343 and U106 in different parts of Europe and the Mediterranean. These can be used to approximate corrections for 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.

How to read this chart

This chart shows how the male-line genetic (phylogenetic) tree splits from its foundation down to the U106 branch. Different current and prehistoric branches are shown in the chart, which can be interpreted carefully.

Where quoted, ages are given as 95% confidence intervals, which we call “2-sigma”. We are 95% sure that the real dates lie between these two boundaries. By dividing the uncertainty 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 2939 BC and 2295 BC.

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

How to read this chart

This chart shows how the male-line genetic (phylogenetic) tree splits from its foundation down to the U106 branch. Different current and prehistoric branches are shown in the chart, which can be interpreted carefully.
Origins of U106 clades: *Homo sapiens* to R1b-M269

**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 Neanderthalic and Denisovan DNA, so one interpretation is that dawn of man occurred at this point (see Karmin et al. 2015 for date estimates).

**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.

**Out of Africa**

Descendants of “Adam” spread throughout Africa. While the Neanderthals 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).

**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.

**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.

**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.)

**Mal’ta**

The origin of R1b 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.

**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 descendies 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.

**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 Dnieper-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.

**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.

**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 descendies 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.

**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.

**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.)

**Mal’ta**

The origin of R1b 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.

**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 descendies 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.

**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 Dnieper-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.

**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.

**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 descendies 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.

**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 Dnieper-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.

**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.
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:

\[ t = \frac{m \cdot r}{n} \]

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 (\(\delta m\)).

For large clades, calculating this uncertainty becomes technically impractical, so we use the Gaussian approximation that the number of mutations we see (\(m\)) is the square root of m, and that the 1-\(\sigma\) (86.3%) uncertainty in m is the square root of m, and that the 95% uncertainty is 1.96 \(\times \sqrt{m}\).

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

\[\delta t = \sqrt{\left[\delta m / m\right]^2 + \left[\delta r / r\right]^2}\]

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:

```
X ABC BC
```

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 of 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:

```
X Y ABC BC
```

Here, B & C share a set of SNPs (X,Y and Z). If only B & C test, we 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
```

we then 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:

\[t = \frac{r(\frac{n}{r} - 0.5)}{2}\]

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

The 95% uncertainty on this is again computed from Poisson statistics, but asymptotes to \(+/- 0.475 \times n\) for large \(n\).
FINAL AGE CALCULATION

The final age is determined from three numbers:
Firstly, from the number of SNPs beneath the target:
\[ t = m \cdot r / n \]  \[\text{[T1]}\]
Secondly, from the number of SNPs between U106 and the target:
\[ t_0 = n(U106) - m_0 \cdot r_0 \]  \[\text{[T2]}\]
where \( n(U106) \) 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 (\( n(U106) \) as follows):
\[ r_0 = m(U106) / r(U106) \]  \[\text{[R2]}\]
Thirdly, from the number of SNPs since the parent clade:
\[ t_0 = n(p) - m_p \cdot r_0 \]  \[\text{[T3]}\]
where \( n(p) \) 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:
\[ t = m \cdot r / n \]  \[\text{[T4]}\]
which then gives the equality:
\[ t = t_0 = t_p \]  \[\text{[T5]}\]
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 (\( m/n > m_p/n_p \)). 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-\( \sigma \) uncertainty value, namely:
\[ \delta t = 1.96 \sqrt{(\delta m/n)^2 + (\delta r/n)^2} \]  \[\text{[ET1]}\]
where \( \delta \) is derived from the literature or case studies. Similarly, the uncertainty in \([T4]\) can be derived as:
\[ \delta t = 1.96 \sqrt{(\delta m/n)^2 + (\delta r/n)^2} \]  \[\text{[ET2]}\]
where \( \delta t_0 \) is given from \([R2]\) by:
\[ \delta t_0 = 1.96 \sqrt{(m(U106) - m_0 \cdot r_0)^2 / r(U106)} \]  \[\text{[R1]}\]
In both cases, \( m \cdot \delta m \) and \( m \cdot \delta r \) are given by the highest and lowest value of \( \lambda \), respectively, for which:
\[ \lambda \cdot e^{\lambda} \cdot \text{Pois}(\lambda, k) > 0.1585 \]  \[\text{[ER2]}\]
\[ \lambda \cdot e^{\lambda} \cdot \text{Pois}(\lambda, k) > 0.8415 \]  \[\text{[ER3]}\]
at 1\% and:
\[ \lambda \cdot e^{\lambda} \cdot \text{Pois}(\lambda, k) > 0.025 \]  \[\text{[ER4]}\]
\[ \lambda \cdot e^{\lambda} \cdot \text{Pois}(\lambda, k) > 0.975 \]  \[\text{[ER5]}\]
at 95% confidence, where \( \text{Pois}(\lambda, k) \) 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 \((\pm \delta t)\) for U106 and the parent SNP, respectively, rather than as an age from the present day.

If \( \delta t < \delta \) (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 \( t_\text{f} \), with associated uncertainty \( \delta t_\text{f} \). Note that because \([ER4]\) and \([ER5]\) do not provide errors symmetric around \( m \), the final uncertainty, \( \delta t_\text{f} \), 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:
\[ \delta t_\text{f} = \frac{w_0 \delta t_0 + w_\text{f} \delta t_\text{f}}{w_0 + w_\text{f}} \]  \[\text{[ET3]}\]
where the weights are set as follows:
\[ w_0 = (t/n - 2\delta t)^2 \]  \[\text{[ET4]}\]
\[ w_\text{f} = (t^* - t)^2 \cdot 2\delta \]  \[\text{[ET5]}\]
where \( t^* \) is either \( t_0 \) or \( n(U106) \), 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 date in the testing population. This comes from an average birth year of these testers is 1950.3 with a standard deviation of 1.50 years for the total BigY testing population). It therefore, it is derived from the literature or case studies. Similarly, the uncertainty in \([T4]\) can be derived as:
\[ \delta t = 1.96 \sqrt{(\delta m/n)^2 + (\delta r/n)^2} \]  \[\text{[ET2]}\]
where \( \delta t_0 \) is given from \([R2]\) by:
\[ \delta t_0 = 1.96 \sqrt{(m(U106) - m_0 \cdot r_0)^2 / r(U106)} \]  \[\text{[R1]}\]
In both cases, \( m \cdot \delta m \) and \( m \cdot \delta r \) are given by the highest and lowest value of \( \lambda \), respectively, for which:
\[ \lambda \cdot e^{\lambda} \cdot \text{Pois}(\lambda, k) > 0.1585 \]  \[\text{[ER2]}\]
\[ \lambda \cdot e^{\lambda} \cdot \text{Pois}(\lambda, k) > 0.8415 \]  \[\text{[ER3]}\]
at 1\% and:
\[ \lambda \cdot e^{\lambda} \cdot \text{Pois}(\lambda, k) > 0.025 \]  \[\text{[ER4]}\]
\[ \lambda \cdot e^{\lambda} \cdot \text{Pois}(\lambda, k) > 0.975 \]  \[\text{[ER5]}\]
at 95% confidence, where \( \text{Pois}(\lambda, k) \) 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.

CHOOSING A MUTATION RATE

We have so far ignored how the choice of the underlying mutation rate, \( r \), and its uncertainty, \( \delta r \), 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.
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 paternaly 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 millennia.

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 neglibible, 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. The infinite allele model we use derives from Dean McGee's work on this.

**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 one "multi-step" mutation, whereas it is possible it was really two mutations: 15→16→17. The step-wise allele model is better.

**Variance-based model:** Both the step-wise and infiniteallele 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/yttility.html](http://www.mymcgee.com/tools/yttility.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.

**Infinite Age Combinations**

![Infinite Age Combinations Diagram](image)

**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 easier to use for STR calculations, which compare two clades to each other.

**Variance-based age calculation**

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

$$\text{Var}(AB) = \frac{s_{ij}^2}{mn} + \frac{s_{ij}^2}{mn} - \frac{1}{m} + \frac{1}{n}$$

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

$$\sigma^2(\text{Var}(AB)) = \frac{\text{Var}(AB)}{\text{Var}(AB)} = 2 \left( \frac{m-1}{m} \right)^2 + 2 \left( \frac{n-1}{n} \right)^2$$

Variances on individual markers can be summed, such that:

$$\text{Var}(AB) = \sum_{i=1}^{m} \text{Var}(AB_i)$$

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

$$\sigma(\text{Var}(AB)) / \text{Var}(AB) = \sqrt{\sum_{i=1}^{m} \sigma^2(\text{Var}(AB_i))} / \text{Var}(AB)$$

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

$$t(AB) = \frac{\text{Var}(AB)}{\sum_{i=1}^{m} \text{Var}(AB_i)} / \mu$$

and:

$$\sigma(t(AB)) = \frac{\sigma(\text{Var}(AB))}{\mu} \sqrt{\sum_{i=1}^{m} \left( \frac{\sigma^2(\text{Var}(AB_i))}{\text{Var}(AB_i)} \right)^2}$$

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

**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 (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 15, 15, 16, 17, 18 or 18, 17, 16, 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.

<table>
<thead>
<tr>
<th>Source</th>
<th># Used</th>
<th>Not Used</th>
<th>Relative Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chandler (2006)</td>
<td>67</td>
<td>50</td>
<td>78%</td>
</tr>
<tr>
<td>Doug McDonald (unpub.)</td>
<td>80</td>
<td>not used</td>
<td>247%</td>
</tr>
<tr>
<td>Charles Kerchner (unpub.)</td>
<td>67</td>
<td>not used</td>
<td>304%</td>
</tr>
<tr>
<td>SMGF</td>
<td>30</td>
<td>not used</td>
<td>109%</td>
</tr>
<tr>
<td>FTDNA</td>
<td>37</td>
<td>not used</td>
<td>268%</td>
</tr>
<tr>
<td>SMGF/Y-Search</td>
<td>21</td>
<td>not used</td>
<td>117%</td>
</tr>
<tr>
<td>Y-HRD</td>
<td>16</td>
<td>not used</td>
<td>83%</td>
</tr>
<tr>
<td>Vermeulen et al. (2009)</td>
<td>8</td>
<td>not used</td>
<td>143%</td>
</tr>
<tr>
<td>Marko Heinila (unpub.)</td>
<td>111</td>
<td>94</td>
<td>78%</td>
</tr>
<tr>
<td>Ballantyne et al. (2010)</td>
<td>91</td>
<td>82</td>
<td>118%</td>
</tr>
<tr>
<td>Burgarella et al. (2011)</td>
<td>84</td>
<td>83</td>
<td>118%</td>
</tr>
</tbody>
</table>

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

Calibration of STR to SNP Ages: Variances

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

\[ f_{\text{STR,corr}} = f_{\text{STR,unc}} \exp(-t_{\text{STR,unc}}/f) \]

where \( f_{\text{STR,corr}} \) and \( f_{\text{STR,unc}} \) 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:

\[ f_{\text{STR,corr}} = f_{\text{STR,unc}} \exp((-t_{\text{STR,unc}}/f)) \]

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

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

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.

We show the following fitting parameters are derived:

\[ f_1 = 4436 +/- 117, \quad f_2 = 4519 +/- 451, \quad f_3 = 0.99 +/- 0.07. \]

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.

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.
U106 family tree

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

INTERPRETING DATES

These dates have been calibrated to existing data using 135, 14 years per SNP. A 95% confidence interval of 115-150 years per SNP has been estimated, hence the absolute ages of such clades are uncertain by +/-125. Thus U106 male line traces to around ~560 years. Additional uncertainties due to random sampling of lines becoming 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 K1, P311. The U106 male L21 cluster spans the expected E1b1b1i1 P311 subgroup of around 780 BC. This date derived from L21 for P311 is 790 BC. The date derived from DF27 is 290 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 R1 is not found in Europe before 2400 BC, despite extensive testing of archaeological remains from prior millennia. 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 (Kromeritz) and England in the period 2800-2600 BC. It is clear from the archaeology that the migration into Europe of the ancestors of U106 happened sometime in the period 2900-2400 BC.

It is reasonable to assume that such an incursion is associated with a population expansion, leading to the production of many new SNPs. The period 3000-2500 BC could be the origin of this expansion is likely to be P311, which greatly dominates all Europe.

The timing of the origin of U106 and P311 can logically be linked to archaeological horizons during this period, hence the expansion of a series of cultures waddling across Europe during the period 3000-2500 BC.

INTERPRETING STRUCTURE

The sizes of individual SNPs refer to numbers of testers sharing those results today and the relative frequency at which people test. For example, it is clear that Z305 was ultimately more successful than its brother clade, Z1748. 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 gleaned from the large gaps and blooms in the phylogenetic record. For example, the large gap between L248 and Z27 and Z28 could be due to a population crash during this period, or simply the dominant population migrating out of an area where it was well tested. Conversely, the large number of lines stemming from Z27 and rapid succession of SNPs between Z27 and Z28 could indicate a 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 expansion of U106 that propagated the two lines displayed in grey. The top structure of L248 and the U106 precursor, S1688. Random differences in the number of SNPs could mean that Z26, Z27 and Z28 formed the tail end of this population expansion, or they could have been chosen at a later date. It is unclear, however, that all major branches then suffered a hiatus in population expansion, or a population contraction around this time, or a part of several lineages is 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 since Z1748, a full millennium later. We can assume that Z18 migrated to an area where it was not able to participate in any major expansions until the post-Roman Migration Age period.

The structure of Z18 appears to have become frozen in shortly after its appearance with no new SNPs since Z1748, a full millennium later. We can assume 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 structure of Z106 does not appear totally random, with limited expansion around 3900 BC (~500 years), 800 BC (~400 years), 400 BC (~400 years), 100 BC (~100 years). This is consistent with Z306 showing a structure evolving later, mainly during the European Bronze Age, indicating a rapid expansion around this time (1500 BC (~600 years)).

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 structure of Z106 does not appear totally random, with limited expansion around 3900 BC (~500 years), 800 BC (~400 years), 400 BC (~400 years), 100 BC (~100 years). This is consistent with Z306 showing a structure evolving later, mainly during the European Bronze Age, indicating a rapid expansion around this time (1500 BC (~600 years)).

Z338 shows a more branching structure, with short branches containing several SNPs (e.g. Z18, Z27, Z2, Z37) with many branches. These are typically followed by long hiatus periods (e.g. Z1749 is S2155, Z2 to Z5).
The geography of U106 testers

Distribution of all U106 by region and sub-clade

This represents the phylogenetic and phylotopic distribution of 157 U106+ tests from Family Tree DNA. The full geographical project information was collected during December 2014 by our members of the U106 project. Geographies is self-reported by the testers. Phylogeographic position is based on the final SNP tested, thus some testers may branch into multiple, ancestral sub-clades.

Distribution of all U106 by European region

Although greatly dominated by British Isles clade P312, Irish U106 is an important component of the expansion, although non-Irish testers are represented in many European regions.

Irish (Republic of and Northern)

29 testers (28% of all Irish)

Although greatly dominated by British Isles clade P312, Irish U106 is an important component of the expansion, although non-Irish testers are represented in many European regions.

The Netherlands shows some interesting comparisons with the British Isles clade P312. In the Irish, there are lower proportions of FGC10799 and L47, but higher proportions of Z30 and Z305. This could be explained by the higher proportion of FGC13326 in the Irish, which is likely to reflect the higher proportion of Irish (republic) origin. The Netherlands shows a higher proportion of Z305, which is likely to reflect the higher proportion of Irish (republic) origin.

Finally, the Netherlands shows a higher proportion of Z305, which is likely to reflect the higher proportion of Irish (republic) origin.

The Netherlands shows some interesting comparisons with the British Isles clade P312. In the Irish, there are lower proportions of FGC10799 and L47, but higher proportions of Z30 and Z305. This could be explained by the higher proportion of FGC13326 in the Irish, which is likely to reflect the higher proportion of Irish (republic) origin.
The geography of U106 testers in the British Isles

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

All of Europe

Northern Isles (Shetland & Orkney)
- Few testers, but probably large Z9 and Z343

England
- 476 testers
- Typical Z9 mix. Notably no U198, but strong Z18 & L47, as in Cornwall.

Scotland
- 422 testers
- Typical Z9 mix. Notably no U198, but strong Z18 & L47, as in Cornwall.

Northern Isles
- 27 testers
- Few testers, but probably large Z9 and Z343.

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 using a “minimal spanning tree”-like method.

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 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 millennium 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 millennium 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.

FURTHER ANCIENT DNA

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

6500-4000 BC: the situation in Europe

<table>
<thead>
<tr>
<th>Location</th>
<th>Haplogroup(s)</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loschbour</td>
<td>I2a1</td>
<td>6220-5990 BC</td>
</tr>
<tr>
<td>La Branya</td>
<td>C1a2</td>
<td>5940-5690 BC</td>
</tr>
<tr>
<td>Motala</td>
<td>I2c2, various I2a1</td>
<td>5898-5531 BC</td>
</tr>
<tr>
<td>Sok River</td>
<td>R1b1-M343, L278</td>
<td>5650-5555 BC</td>
</tr>
<tr>
<td>Yuzhnyy Oleni Ostrov</td>
<td>R1a1-M459, Pages65.2</td>
<td>5500-5000 BC</td>
</tr>
<tr>
<td>Alsónyék-Bátaszék &amp; Lánycsók</td>
<td>2<em>F (GHIJK?), H2, G2, 3</em>G2a, G2a2b, I</td>
<td>various dates between 5840-5550 BC</td>
</tr>
<tr>
<td>Tiszaszölös-Domaha´za</td>
<td>I2a</td>
<td>5780-5650 BC</td>
</tr>
<tr>
<td>Vinkovci &amp; Vukovar</td>
<td>G2a, G2a, I2a1</td>
<td>est. 6000-5500 BC</td>
</tr>
<tr>
<td>Els Troncs</td>
<td>F (xG, …), I2a1b1, R1b1-M343(xM269)</td>
<td>5311-5068 BC</td>
</tr>
<tr>
<td>Tolna-Mözs</td>
<td>G2a2b, F-M89 (xGHIJK?)</td>
<td>various dates between 5300-5020 BC</td>
</tr>
<tr>
<td>Berekalja</td>
<td>C1a2</td>
<td>5300-4950 BC</td>
</tr>
<tr>
<td>Derenburg</td>
<td>2*F-M89 (xGHIJK), G2a2b</td>
<td>5300 - c. 5000 BC</td>
</tr>
<tr>
<td>Halberstadt</td>
<td>3<em>G2a2a; 2</em>G2a2a1</td>
<td>various: 5207-4946 BC</td>
</tr>
<tr>
<td>Kompolt-Kigyoser</td>
<td>C1a2</td>
<td>5210-4990 BC</td>
</tr>
<tr>
<td>Szölőskert-Tangazdaság</td>
<td>G2a2b</td>
<td>est. 5100 BC</td>
</tr>
<tr>
<td>Avellaner</td>
<td>3*G2a, E1b-M35.1+ V13+</td>
<td>5000 BC</td>
</tr>
<tr>
<td>Lengyel</td>
<td>R1a R1b</td>
<td>4490-4360 BC</td>
</tr>
<tr>
<td>Balatonszemes-Bagódomb</td>
<td>I1-M253</td>
<td>est. 5000 BC</td>
</tr>
<tr>
<td>Avellaner</td>
<td>3*G2a, E1b-M35.1+ V13+</td>
<td>5000 BC</td>
</tr>
</tbody>
</table>

Iberia:

- 30% G
- 20% I
- 10% C
- 10% E
- 10% H
- 10% R

Central Europe:

- 58% G
- 21% I
- 8% C
- 4% T, H, E

Russia:

- 100% R

Scandinavia:

- 100% I
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.

4000-3000 BC: the rise of M269

Russia: 100% R

Central Europe:

83% G

17% I

Western Europe:

100% G

20% I

100% R

83% G

17% I
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.

2600 BC (3050 BC - 2100 BC)
2700 BC (3450 BC - 2150 BC)
3000 BC (4450 BC - 2800 BC)
4000 BC (5000 BC - 3100 BC)
4100 BC (5350 BC - 3200 BC)

3000-2500 BC: M269 invades Europe

Ukraine:
67% R
33% I

Southern & Western Europe:
77% G
23% I

Central Europe:
75% R
8% G?
17% I (or J?)

Southern Scandinavia:
67% R
33% I

Russia:
83% R
8% N
8% I

Scandinavia:
67% R
33% I

Western Europe:
77% G
23% I

Southern:
67% R
33% I

Northern:
75% R
8% G?
17% I (or J?)

Smaller:
83% R
8% N
8% I
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.

---

**Central Europe:**
- **54% R**
- **46% I**

**Russia:**
- **67% R**
- **16% N**
- **16% I**

**Southern Scandinavia:**
- **75% R**
- **25% I**

---

2500-2000 BC: early growth of U106

**Updated: 17 June 2015**

Dr. Iain McDonald
on behalf of the U106/S21 group
The origins of U106: 3800 BC to 2650 BC

1. Arrival into Europe
   The origin of U106 can now be placed around 2650 BC. There is roughly a 2–3:1 chance of it being within 250 years of this date. Ancient DNA shows the haplogroup R1b belongs to Europe before about 3000 BC, so we can presume that U106 was founded somewhere 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 larger scale of the Kurgan expansion taking place at this time fits the Kurgan expansion of the Russian Steppes.

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 geographical and the east–west trend. 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 Gimbatura interpretation of the Kurgan hypothesis credits the Kurgans with the introduction of the Indo-European 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 Caucasus to the trans-Ural region shown here.

5. M269
   M269 appeared around 4000 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 Donbas, took place around 4500–4000 BC and could be an early origin for M269. Wave 2 probably occurred from the Mycropolis culture (3700–3600 BC), indicated in the high. The variant 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 Oriental, Caucasian, Turk–Mongol and European M269 populations. L51 does not clearly appear in the ancient DNA results itself. However, ancient DNA from the Trypillia culture (modern-day Ukraine) shows that this and its subclade Z2013 dominate here. ADDitionally, FTDNA shows M269+ L23+ and L23–L51 tests are more focused 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 past PF7500 and its subclades are clearly concentrated in the Balkans and eastern Mediterranean, CT5782, particularly CT59219, is closely associated with the Balkan peninsula. L272 is spread around eastern Europe; while CTS7865 may be Balkan or Anatolian. Meanwhile, ancient DNA from Hauk et al. (2015) shows a very strong expansion of Z2013 throughout the 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 PF7500. L51 may be mostly L23→Z2013. Combined with the Hovhannisyan population results, it suggests they are probably mostly L23→Z2013.

9. M269→L23→PF7500 and the Hittites
   The variance of PF7500 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 PF7500 to a Hittite population. 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 geographical and the east–west trend. In addition, it appears to arise from the trans-Ural area near concentrations of groups further up the haplogroup R1b tree (e.g. V88).

10. L51→PF7500: migration into Europe
    L51 through to PF7500 represents an 800-year gap in our knowledge. This may represent a hiatus in the east–west population expansion 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→PF7500 should represent the European focus of L51→Z2013, but is just widely found in Germany. The median location of PF7500 in FTDNA tests is close to the Black Sea. The east–west migration means the split between P311 and P312 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. M269→L23→L51
   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 geographical and the east–west trend. In addition, it appears to arise from the trans-Ural area near concentrations of groups further up the haplogroup R1b tree (e.g. V88).

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 slightly earlier, and considerably further east.

13. 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 P1a1 dominates) suggests it cannot have been further east than the German–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.

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 found this foundation at around 2850 BC, give or take a few centuries, which allows U106 to form both the Single Grave and Bell Beaker cultures, plus the proto-Celtic cultures that followed them.

The U106 ancient DNA from Líttle Beddingham in Sweden (sketch RISE98) show that our ancestors were ‘Idile, 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 ideal 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 TMRCA will be a generation younger. Of each of those half, 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:

- G106>2381>215606998
- G106>2381>2301X>480>29>23>5798

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:

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.

CHARACTERISING RANDOM SPREAD THROUGH INTER-CLADE TMRCA DISTRIBUTIONS

We can now take our example clades, DF98 and Z8, and measure their characteristic infinite 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 TMRACAs for DF98-Z8 pairs, we arrive at the following histogram (binned into 120-year bins):

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 histograms 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 Z38>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 TMRCAs, 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/t law.

These will each be smeared out with a Gaussian of 200–300 years in width in uncorrected 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.

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 TMRCAs 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 considerably younger than the original population. The use of genetic variance as an indicator of origin is a well-known tool. However, statistical spread can cause substantial differences on its own, so care must again be taken when using this method.
**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).

**Tentative migration map**

NB: Do not take this map as proven fact!

**Factsheet: U106**

Dr. Iain McDonald on behalf of the U106/521 group

**Infinite alleles TMRCA**

**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).