U106 explored: its relationships, geography and history
Report to the U106 group, Mar 2016 edition
Principal investigator: Iain McDonald

Contents

Background to genetic testing
Foreword & basics of Y-DNA testing
General testing advice
Overview of the Y chromosome
Characterisation and quality control of next-generation tests

History before U106
Deep ancestry of U106: 200 000 BC to 2500 BC
The route out of Africa to M269 (200 000 BC to 4000 BC)

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

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

Spread and ancestry of U106
Methods for determining the source, direction and date of migrations
Evidence from archaeological DNA
Hypothesised migration of U106 ancestors into Europe
The question is: is it wrong for it to be useful.

The dates I compute here were outdated before I finished compiling the information.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

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 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 alternately S21, and is the result of a simple typographical error that 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

TABLE 2: DYS1234 Mutations

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>U106</td>
<td>3%</td>
</tr>
</tbody>
</table>

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. Thanks are also due to the rest of the U106 team - primarily Charles Moore and Raymond Wing - and David Carlisle and Andrew Booth for sorting the details out.

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.
General testing advice

Everyone's situation is different. The best testing route depends on your budget, on your DNA matches, on their budgets (or your ability/willingness to pay for them), on what you are hoping to get out at the end, and what you can realistically achieve given the limits of money, people and technology. Everyone's journey is different.

What do you want to do with DNA testing? Find the origin of your immigrant ancestor? The origin of their surname? Find out when and how their ancestors arrived in Britain (or any other country)? Or perhaps find out what your deep prehistoric roots are? I'll address these points below, but please bear in mind that this testing advice is general: use your judgement to determine whether this applies to your particular situation.

The testing advice to most people is fairly similar: maximise what you can find out about your own DNA, then carefully select people around you to upgrade – either at their expense or yours. This strategy stems from the basic principle that DNA is a comparative science: your results only mean something if you have someone else to compare them to. You will make the best progress by taking charge and directly engaging with the people to whom you are most closely related. The following testing advice is written from the point of view of U106, but it can readily be translated to any other haplogroup.

PREHISTORIC AND EARLY HISTORIC ROOTS

Are you Celtic or German, or even Saxon, Norman, Viking, Flemish, Angle, Jute, or maybe something else? Obviously, your Y-DNA is only a very small part of that story. Even only 600 years ago, your Y-DNA will account for less than 0.001% of your ancestry. But we are beginning to unravel the ancient roots for many Y-DNA groups. The Family Tree DNA haplogroup projects operate at the interface between amateur genetic genealogy and professional genetic anthropology. Both of us are trying to uncover the (pre-)historical migrations of the world: the professionals focus on the large-scale structure, whereas our interests are typically more specialised.

Step 1: SNP testing. The process of discovering your prehistoric ethnic Y-DNA ancestry relies on grouping together people with the same SNP mutations to form new branches of the human haplotree. The best thing you can do here is take a next-generation test like BigY or YElite to uncover the SNPs in your line. If you can't afford these tests, test the appropriate SNP pack(s) at Family Tree DNA or YSeq, or take a Chromo2 test. YSeq is generally cheaper: this is especially true for people who are probably or confirmed to be U106+, but who don't know which part of U106 they belong to. The U106-L48 pack covers all of U106, and only one pack needs testing, rather than two. For tests with companies other than Family Tree DNA, you should report your results to your haplogroup administrator(s). You should only order an SNP pack if you do not intend to take a next-generation test, or you will be paying for the same thing twice.

For anyone in R-M269, any of these tests will probably take you what is happening in the period between 5000 years ago and about 2000 years ago. Some people are populated by SNPs that are less than 1000 years old; some people are stuck in rare clades that are over 4000 years old. Either way, you will find yourself related to one or more people on your "terminal" SNP. Your goal is to find people who are related to you more closely.

(* Terminal is a bad word to use here, but common parlance. It refers to the most-recent SNP you share with someone else. You will also find a lot of SNPs that are discovered in only your test, which we term "singleton". You want to find people who share some of these singletons with you.)

Step 2: fully upgrade your STR markers, and find your close matches. To find people you need close matches. Most people have only taken STR tests, not SNP tests. If you have not done so already, upgrade to 67 or (preferably) 111 STR markers. This will help you identify the maximum number of matches. Don't worry if you don't match anyone at 37 or 67 markers, you will match someone somewhere, and the more markers you test, the more we can beat down the random noise in the mutations to see who you match. Matches beyond the Family Tree DNA system can be found using YSearch, or Semargl.me.

If you feel capable, try to identify the STR mutations you have in common from an older modal (e.g. the U106 modal). You can then use this template on YSearch.org, semargl.me or simply the table of results from Family Tree DNA projects, to identify those people who also share some or all of these mutations. If you aren't happy doing this, identify the genetic distances of the people who are positive for your "terminal" SNP (or upstream SNP if you have none), then look for people who have not tested SNPs who match you with a smaller genetic distance.

Step 3: encourage your close matches to upgrade. If you have taken a next-generation test (BigY/YElite/etc.), you should also ensure your closest match has upgraded to the same number of STR markers as you. If you have taken an SNP pack, encourage your STR matches to take the same SNP pack or test your individual "terminal" SNP. From hereon, I shall assume you have taken a next-generation test, as the following steps otherwise don't make much sense.

You should then encourage your close STR matches to upgrade to BigY or YElite. They will hopefully share some of your singletons, and give you a new "terminal" SNP. If you can't cajole them into taking these tests, or pay for them, then you should encourage them to test your singleton SNPs at YSeq.net. You should particularly concentrate on those people who have European ancestry, as they can tell you where your "terminal" SNP originated. Devote even more energy to those from continental Europe, where our coverage is poorer, and especially places like France and eastern Europe where legal or economical reasons mean we have very few testers.

In this way, you will gradually bring what is known about your history forwards towards the present day. How far forward this will bring you depends on the number of matches you have, and how willing they are to upgrade.

HOW DID YOUR ANCESTORS ARRIVE IN BRITAIN?

The testing advice for this is goal pretty much the same as that above. Crucially, however, you need to target the period when your ancestors are likely to arrive in Britain. This will depend on your haplogroup. For most U106, particularly those in Z8, it will typically be in the range 800-2100 years.

You need to start by working out your closest STR and SNP matches from continental Europe. You need to find phylogenetically when you were last related, by testing yourself and them with either a next-generation sequencing test (BigY/YElite) or SNP packs, and get them to upgrade their STR results to match your own if necessary. This gives you an upper limit for the length of time your family has been in Britain, though it is only valid if your continental matches did not migrate back to Europe. This seems comparatively rare but obviously did happen.

Generally speaking, we don’t yet have enough data to determine that SNPs are specifically British, although there are a few recent cases (within last 1000 years or so for U106) where we can be quite sure. The biases in our samples mean that it only takes one distinctly related tested to cast significant doubt on the “Britishness” of any SNP. For U106, this is something we are actively researching as a group, where can apply statistical techniques across U106 to say groups are likely to be related on a particular timescale.

RECENT RELATIONSHIPS: SURNAME ORIGINS & ANCESTORS OF AMERICAN IMMIGRANTS

These problems are very much related (no pun intended). If you want to origins this recent, you obviously have one more piece of information: your surname. This means you can think a bit differently. You might also want to consider anyone who is expected to match you within the last 1000 years (refer to comparison tools like http://www.mymgcce.com/tools/yutility111.html or ages estimates like those from YFull or (for U106) my analyses). I will presume that your ancestors originate from the British Isles, but the same basic principles can be applied to other locations.

Step 1: upgrade your own DNA. The first thing to do is see what you can learn from your own DNA. Upgrade your STR markers to a reasonable level (67 or 111). This will let you accurately see who matches at lower levels can be spurious, unless you are lucky enough to have significant numbers of historical STR mutations in your first 37 markers. It is also impossible to get a sufficiently accurate TMRCA (time to most-recent common ancestor) when you were last related, by testing yourself and them with either one or more people on your "terminal" SNP. Your goal is to find relationships will be uncertain by about a factor of two.

You need to start by working out your closest STR and SNP matches from continental Europe. You need to find phylogenetically when you were last related, by testing yourself and them with either a next-generation sequencing test (BigY/YElite) or SNP packs, and get them to upgrade their STR results to match your own if necessary. This gives you an upper limit for the length of time your family has been in Britain, though it is only valid if your continental matches did not migrate back to Europe. This seems comparatively rare but obviously did happen.

Generally speaking, we don’t yet have enough data to determine that SNPs are specifically British, although there are a few recent cases (within last 1000 years or so for U106) where we can be quite sure. The biases in our samples mean that it only takes one distinctly related tested to cast significant doubt on the “Britishness” of any SNP. For U106, this is something we are actively researching as a group, where can apply statistical techniques across U106 to say groups are likely to be related on a particular timescale.

RECENT RELATIONSHIPS: SURNAME ORIGINS & ANCESTORS OF AMERICAN IMMIGRANTS

These problems are very much related (no pun intended). If you want to origins this recent, you obviously have one more piece of information: your surname. This means you can think a bit differently. You might also want to consider anyone who is expected to match you within the last 1000 years (refer to comparison tools like http://www.mymgcce.com/tools/yutility111.html or ages estimates like those from YFull or (for U106) my analyses). I will presume that your ancestors originate from the British Isles, but the same basic principles can be applied to other locations.

Step 1: upgrade your own DNA. The first thing to do is see what you can learn from your own DNA. Upgrade your STR markers to a reasonable level (67 or 111). This will let you accurately see who matches at lower levels can be spurious, unless you are lucky enough to have significant numbers of historical STR mutations in your first 37 markers. It is also impossible to get a sufficiently accurate TMRCA (time to most-recent common ancestor) when you were last related, by testing yourself and them with either one or more people on your "terminal" SNP. Your goal is to find
While these SNP tests are useful, they are generally less important. If you have a good idea of where to start, you will need to interpret a series of otherwise meaningless numbers. Most of the time, you will have to go and encourage people to take up DNA tests which will let you see if people from these regions as best you can, and identify lines to research. However, they will still require a modest technical knowledge to interpret interpretation on top of that. These results are better quality controlled. In many ways, this is similar to receiving your STR results for the first time. You will need to interpret a series of otherwise meaningless numbers.

The next step is to get DNA from people in these hot spots. If you are fortunate enough to have existing matches at 37, 67 or 111 markers who share your surname and who either still live in Britain or can trace their ancestry back to a particular place in Britain. Sites like Ancestry and Genealogy websites are useful for identifying such people. If you then need them to test their Y-DNA, and convincing them to do so means you need to formulate the cheapest, most reliable test for them. This will depend on whether you have any mutations that uniquely identify your family in the first few markers. If so, a 12-marker test and a single SNP test might be the cheapest way. Otherwise a 37-marker test might be preferable. (NB: you now have to ask for 12-marker tests specially.) You could even just get them to test a single SNP test for your "terminal" SNP at Yseq.net, but this will simply give you a binary indicator of whether or not they are related to you. A wider test will give them some extra information about their origins regardless of whether they are related to you or not. Remember that if they are paying, they are relying on you to provide good advice, so remember to check in their interests as well as your own!

**Step 3: identifying likely geographical origins.** The next step will pin down the geography of your surname. This is the most difficult, and perhaps the most expensive bit. Your success, and the order you do things in, depends on the frequency of your surname, whether you can find it in old historical records (which may by topographic in origin), the number of existing matches you have, and how many you can convince to pay for themselves. The rest of this is therefore even more generalised advice.

If you genuinely have no idea where in Britain to start, search for your surname here: [http://gbnames.publicprofiler.org/Surnames.aspx](http://gbnames.publicprofiler.org/Surnames.aspx) and identify the regions where it is most common. If there is only one region, that gives you a good idea of where to start. If not, you will have to pick off these regions one by one.

Once you have found a region, you need to narrow things down by parish. You can also use resources like FreeCen, the IGI and Scotland's People to map the number of people sharing your surname by parish. On some pay-per-view sites like Scotland's People, searching is still free, so it is still possible, if laborious, to obtain numbers. Old historical records can be very useful too, such as pre-1841 tax registers. Mapping the number of people with your surname (or its variants) in these will typically show one or more "hot spots" on a parish level that you can work with further.

Common surnames pose more of a problem. A case study in this is my own historic surname Donald, which is quite common in Scotland. There are at least 11 distinct origins for the Donald surname in Scotland. Mapping all of Scotland at the parish level has also identified at least 11 distinct "hot spots". My family traces to Aberdeen, as does that of my closest non-Donald match. Even within a few miles of Aberdeen, there are three distinct geographical groups that show up, that resolve into at least four different origins for the surname. In such cases, a lot of testing needs to be done to identify the relationship between geographic and genetic groups.

**Step 4: probing “hot spots”.** The next step is to get DNA from people in these hot spots. If you are fortunate, some people will already have had a DNA test, and your project administrator may be able to provide you with information on how their origins and lines are related to you or not. If you are unlucky, which will be most of the time, you will have to go and encourage people to take up DNA testing who haven’t done already.

A good source of such people is the wider genealogy community. The important thing is to target people who have long, secure and unbroken lineages who share your family surname and who either still live in Britain or can trace their ancestry back to a particular place in Britain. Sites like Ancestry and genealogy websites are useful for identifying such people.

You will then need them to test their Y-DNA, and convincing them to do so means you need to formulate the cheapest, most reliable test for them. This will depend on whether you have any mutations that uniquely identify your family in the first few markers. If you do, a 12-marker test and a single SNP test might be the cheapest way. Otherwise a 37-marker test might be preferable. (NB: you now have to ask for 12-marker tests specially.) You could even just get them to test a single SNP test for your "terminal" SNP at Yseq.net, but this will simply give you a binary indicator of whether or not they are related to you. A wider test will give them some extra information about their origins regardless of whether they are related to you or not. Remember that if they are paying, they are relying on you to provide good advice, so remember to check in their interests as well as your own!

**Step 5: detailed exploration.** Once you have covered all the "hot spots", you should find out where people are related to you, so roughly where your ancestors came from. You should then focus on this region, and get a few people with your surname to test (and surnames of your closest STR matches if applicable).

This will give you some idea of the breadth of variation in STR types, and which STR mutations happened when in your family. You should then construct a phylogenetic family tree, showing how your lines fit together and when each mutation is likely to have occurred. Note that this may require SNP testing to accomplish: first a next-generation test (BigY/YElite) to identify novel variants, then single SNP testing at Yseq.net to identify who you share those variants with. Try to reconstruct family trees of everyone with your surname from these regions as best you can, and identify lines to research further. Very few people can expect to extend their paper trail via this method, but it should give you a good idea of where your ancestors came from and some key local figures to whom you may be related. This will then allow you to proceed any ancestry project from within about 10 miles of the City of Aberdeen, where it has been since at least 1400 AD.

Once again, everyone's situation is different. One of the things that haplogroup projects, like the U106 group, are best at is individual recommendations for testing. However, it's likely that it will be based around this kind of procedure. If you are considering further testing at this point, see if these situations can be adapted to what you want to accomplish, and your haplogroup project should be able to you more individualised advice.

### BIGY OR YELITE?

The choice of next-generation sequencing is not so clear cut in today’s market. There is not a one-size-fits-all solution to everyone’s problem. What you do depends on your particular problem and how best it is solved.

**Coverage:** The Y chromosome is about 60 million base pairs long. BigY sequences 8-10 million of these (see later pages). YElite sequences about 14 million of these. Read quality is generally better for YElite than BigY: the number of SNPs found in a YElite test is 40-60% greater than in BigY. Neither test covers every important SNP, but both cover most of them. E.g., in U106, YElite covers Z301 and DF98 but BigY doesn’t. Generally speaking we will know whether you are positive for most of these kind of SNPs, but some people will be exceptions.

**Value for money:** The FGC/YFull test is about 55% greater than BigY ($775 versus $575: both prices are subject to change and discount). Since YElite has greater coverage, it is better in terms of SNPs per dollar.

**Analysis:** BigY will give you a list of known SNPs, which include most of those on their haplotype and plenty more besides. They will also give you a list of "novel variants" – some of these are SNPs we’ve known about for some time, some will be new to your test, and some will be bad data (no test is perfect). Almost everyone will need some help in analysing the results from BigY beyond what FTDNA will give you. Your haplogroup project (e.g. the U106 group) can help.

YElite gives you a set of raw results, with limited interpretation on top of that. These results are better quality controlled. However, they will still require a modest technical knowledge to understand. I would anticipate that most people will need help in analysing the results. Again, your haplogroup project can help.

In many ways, this is similar to receiving your STR results for the first time. You will need to interpret a series of otherwise meaningless numbers.

**Analysis by a group:** Different groups have different ways of dealing with data. In the U106 group, we are well set up to rapidly analyse the results of BigY tests and report back to individuals. We can tell you how and when you relate to other testers. We are starting to offer suggested origins in a few cases where these are becoming clear.

In U106, we are not yet at that stage with YElite. Individually, I am not yet able to provide ages for YElite tests, as the systems are sufficiently different that I need to characterise the test results and work out how the differences between the different WGS/YElite tests affect the age estimates. This is work in progress: we intend to have something working soon.

**Incorporation of data:** YElite tests obviously will not be incorporated into the Family Tree DNA database. You will have to keep track of more things yourself, e.g., how your haplogroup relates to those of the people around you, and making sure all your project administrators know you have YElite etc. It’s a bit more work, but one to be considered.

For U106, our Yahoo forum acts as a repository for this information, where we can process everyone's tests together.

Conversely, FGC will name your SNPs and officially "register" them, which FTDNA is not yet doing. If you have a BigY test, you can pay FGC/YFull $49 for this privilege. FGC will also put you into their tree and incorporate into the Family Tree DNA database. You will have to keep your own project administrators informed about your results. Most people will have an FGC test at some stage, so this is useful.

Another advantage of YElite is that anyone can have one expensive SNP test, and the eventual cost of later steps.

However, they will still require a modest technical knowledge to understand. I would anticipate that most people will need help in analysing the results. Again, your haplogroup project can help.
THE Y CHROMOSOME

The Y chromosome is the shorter of the human sex specific chromosomes. The reference sequence we use in this document, Build 37, stretches it out to 59,375,566 base pairs. (The newer Build 38 is slightly shorter, as several gaps have been closed.)

When a genome is sequenced, the DNA is broken up into bite-sized pieces of tens or hundreds of base pairs long. These pieces contain genetic code in the form GATACTGA… They run like stretches of tape. The reference sequence is a bunch of these, stitched together where they overlap. However, there are gaps in this sequence, so we do not even know exactly how long the Y chromosome is, never mind what is in it.

In reconstructing a genome from a new test (e.g. BigY@Yel), fragments are compared to this reference sequence, and pasted in where they best fit. New SNPs are discovered by looking for differences from the reference sequence.

This process requires knowing: (a) which chromosome you are looking at and (b) which place on the chromosome you are looking at, so regions that look like other chromosomes and very repetitive regions usually cannot be read accurately. Only the euchromatic regions can.

Overview of the Y chromosome

Compiled by: Dr. Iain McDonald; updated: 2 Dec 2015

Source material:
P. Francalacci, et al., 2013, Science, 341, 565

Selected loci from: ybrowse.org

Selected text from: ycombinator.org

Selected text from: go2genomics.com

Selected text from: Y-Elite.com

Selected text from: Family Tree DNA

Selected text from: Full Genome Corp.

Selected text from: DNA Reptile

Selected text from: DNA Interactive
**NEXT-GENERATION TESTING**

“Next-generation” tests like Family Tree DNA's BigY and Full Genome Company's Y-Prime and Y-Elite products offer an unparalleled chance to uncover new SNP mutations, insertions and deletions (indels) in your DNA. These are the only reliable tool we have for determining new structures within the Y-DNA tree (clades) and the only accurate way of obtaining dates we have. Of the 59 million base pairs in the human Y chromosome, these tests only cover between about 8 and 14 million.

**CLADE IDENTIFICATION**

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

- **chrY 2660548 2665410**

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

- **chrY 2661800 2662940**

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

- **22216800 22251940**

SNPs that are called in both tests are marked as consistent (i.e., +). SNPs that are called in only one test are marked as inconsistent (i.e., *).

**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 510 BigY tests: the entire analysed sample as of 17 November 2015. The average BigY test comprises of 10,613,667 callable base pairs (standard deviation 312,666) over 11,358 regions (st.dev. 2391). Typically 131 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, including dating of SNP ages, I have removed the entire DYZ19 region between positions 22216800 and 22251940 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.

The typical overlap between two tests (excluding DYZ19) is 10,504,207 base pairs (st.dev. 308,546), or 97.5% overlap. For two given tests, 2.5% of SNPs are not called in the matching test.

**CHARACTERISATION OF FGC Y-ELITE**

Eight Full Genomes Corp. YElite 1.0 tests were analysed by Vince Tilroe during December 2015. YElite 2.0 tests were not available for comparison at this time, but reports from early batches indicate a similar number of callable base pairs for the YElite 1.0, 2.0 and 30x Full Genome tests.

The average YElite test comprises of 14,073,254 callable base pairs (standard deviation 178,471). This gives it considerably higher coverage than BigY (33% more, according to the respective companies’ coverage — this is explored more later). It also gives it much higher repeatability between tests.

Rather than a simple yes/no flag, FGC assigns quality codes to each SNP: +, *, ** and *** in decreasing order of quality. Taking an average among the 27 existing YElite 1.0 tests registered with the U106 group, there are an average of 559 SNPs of all qualities called in each test. Of these, 170 are common to all U106 testers. Typically each test will have 29.9 shared SNPs below U106 (at all quality levels), and 34.0 singletons will be called at the (+ and * quality levels), making 63.9 SNPs below the U106 level.

At face value, this gives 1.75x more SNPs under U106 than BigY. Note that this is obviously inconsistent with the 1.33x greater coverage. Some of the difference will be because not all the inconsistent SNPs have been found in YElite, as not enough tests have been taken at this stage. Some of the difference will be due to intrinsic differences in the rate of SNP formation in different regions of the Y chromosome. However, a more significant difference will be due to differences in what each company claims to be a callable base pair.

**CHARACTERISATION OF BIGY USING THE FGC PIPELINE**

Vince Tilroe also characterised the raw data (BAM files) of seven BigY tests using the Full Genomes Corp. data reduction pipeline. This provides a consistent basis on which to assign coverage between the YElite and BigY tests, allowing for direct comparison between the two. The results by Y-chromosome region are presented in the chart on the next page.

This analysis highlighted a much reduced number of callable base pairs using the FGC pipeline than the reported coverage by Family Tree DNA. In the following pages, I discuss the differences between the Family Tree DNA and FGC coverage statistics of the BigY and determine which is more appropriate for assigning the number of base pairs called in any particular test.
As mentioned in the previous sections, notable differences exist between the coverage reported by Family Tree DNA of BigY tests, and that reported from analysis of the raw data by a third party, Full Genomes Corp. To investigate which coverage statistics are more accurate, we take each firm’s coverage with the number of repeatable SNPs in that region. These repeatable SNPs are selected and defined by the following steps: (1) they are called by Family Tree DNA in the test results of our 516 BigY testers (the full sample at the time of this analysis); (2) they must be called “PASSED” in the Family Tree DNA variant call file (VCF); (3) they must be found in at least two closely related tests to demonstrate they are accurately called; and (4) they must not be called inconsistently with the rest of the tree phylogeny, in order to ensure that false positives are removed.

The more accurate the coverage, the closer the relation we can expect between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called. We take the square of the Pearson correlation coefficient to determine the expected between coverage and number of SNPs called.

The exclusion of BigY regions is shown below: For all SNP calls that were made, we can count the number of base pairs accurately called in a BigY test, we must determine in which regions the coverage claimed by Family Tree DNA is accurate.
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 male lines. This is a rapidly evolving field, thus the information is subject to considerable change over time.

This tree summarises the entire family tree that lies above U106. This shows how U106, which now represents many tens of millions of men worldwide, branched off from the root of the human Y-chromosome tree at different points in prehistory. A map of this tree is shown on the next page.

(2) OUT OF AFRICA

Ultimately, we all descend from the first life-forms, which lived approximately three billion years ago. Through a long and complicated process, they evolved into hominids. While Homo 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 living person today. We call this person Y-chromosomal Adam, because we all descend 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.

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

The human genetic tree continued to diversify and fluctuate 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 (Ust-Ishim), from a man who was haplogroup R, 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 archeaological remains of Mal’ta Boy, who was buried 24,000 years ago in south-eastern Germany.

(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. R1b is strongest in eastern populations, where it can exceed 60% of individuals in Poland and the south-west Russian states. Its British content is thought to be strongly Viking in origin.

R1a (M343) is thought to have arisen less than 18,000 years ago. In Europe, it is very much dominated by R1a1a2, 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 Asia or South-Eastern Europe.

The date of this expansion into Europe can probably be tied to the sudden growth in the number of branches below M269, which can be very roughly dated to around 4000 BC. The origin of this migration and its route into Europe are not well determined at present. However, archeological remains show that there was extremely few haplogroup R men in Europe before 22,000 BP, when remains from both R1a and R1b are found in Corded Ware and Bell Beaker burials (respectively) in south-eastern Germany.

(4) FOUNDING A NEW EUROPEAN POPULATION

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

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

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

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

The history in this information comes from a variety of sources, but I am most interested to the Social Geographical Genealogical (ISOGG) for maintaining the underlying tree structure displayed here. The anthologue of haplogroup statistics on eupedia.com has also been instrumental in creating these data.

Created by: Dr. Iain MacDonald, updated: 17 Nov 2014

How to read this chart

This chart shows how the male-line (phylogenetic) tree splits from its foundation down to the U106 branch. Different branches of the phylogenetic tree are shown as vertical branches on the chart, whereas branches on the horizontal axis are interpreted carefully.

Where quoted, ages are given as 95.5% confidence intervals, what we call “2-sigma.” We are 95.5% sure that the real dates lie between these two boundaries. By dividing the uncertainty in half, we can recover the 68% confidence interval, or “1-sigma.” For example, the age given above indicates that we are 68% sure that the U106 founder lived between 3260 BC and 1974 BC. We are 68% sure that he lived between 2932 and 2295 BC.

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

Deep ancestry of U106

Acknowledgements

The following data give the number and percentage of various levels between R1-M343 and U106 in different parts of Europe, as found by Myers et al. (2007) and selected other studies. These can be used to approximate correction factors to debias our statistics according to how many men of different ancestries have tested. These numbers are only very approximate in many cases and only represent first-order estimates of the underlying population.

Haplogroup Frequencies in Europe

This chart shows the male-line (phylogenetic) tree splits from its foundation down to the U106 branch. Different branches of the phylogenetic tree are shown as vertical branches on the chart, whereas branches on the horizontal axis are interpreted carefully.

Where quoted, ages are given as 95.5% confidence intervals, what we call “2-sigma.” We are 95.5% sure that the real dates lie between these two boundaries. By dividing the uncertainty in half, we can recover the 68% confidence interval, or “1-sigma.” For example, the age given above indicates that we are 68% sure that the U106 founder lived between 3260 BC and 1974 BC. We are 68% sure that he lived between 2932 and 2295 BC.

This date was calculated using SNP-counting methods which are detailed on later pages.
 Origins of U106 clades: Homo sapiens to R1b-M269

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

(2) Y-chromosomal Adam
In this study, we only trace male lineages. These converge in a more recent ancestor, who we call “Adam”. Our most distant relations are a group of Kalahari, who share the most distant haplogroup, A00. “Adam”. Our most distant relations are a group of Kalahari, who share the most distant haplogroup, A00.

(3) Out of Africa
Descendants of “Adam” spread throughout Africa. While the Neantherthals 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 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.

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

(5) North and East
Haplogroup K represents a large section of the descendent of the people that left Africa. It includes many Amerindian, Australasian, far Eastern and polar populations.

(6) Ust'-Ishim
The origin of haplogroup K is debatable. Several scholars place it along a continuous progression across modern Iran, skirting the Tian Shan mountains towards Lake Biakal, though origins from the Caucasus to south-east Asia have been discussed. Recent hypothetical reconstructions have put the origin of haplogroup K as far south and east as the Indian sub-continent. 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 Fa et al. 2014 for the Ust'-Ishim burial.)

(7) Mal’ta & haplogroup R
The following few millennia are even more difficult to piece together. It appears that our ancestors survived in the Siberian tundra for many tens of thousands of years. An important split, between the Q and R haplogroups, occurred around 250000 years ago. Haplogroup Q went east to east Asia and Siberia. Haplogroup R is the lineage of our ancestors. The exact date isn’t known, but an early haplogroup R burial has recently been sequenced in the region of Mal’ta, just west of Lake Biakal, which dates to 24000 years ago. This burial is probably less than 1000 years before the last common ancestor of all haplogroup R today.

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

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

(10) M269
The story of R1b beyond the Urals is poorly known. What is clear is that there was a gradual movement westwards, probably to somewhere in the Dniester-Don valley system, where it probably arrived in the post-glacial Holocene era. M269 defines the next major split in our ancestral population. It occurred around 6000 years ago, and has been linked to the Yamnaya culture and the expansion of Indo-European languages and culture to Europe around 5000 years ago (the Kurgan hypothesis). 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. The details of the movements between M269 and U106 are explored later in this document.

(11) Ust’-Ishim
denovo SNPs, M343, probably arose less than 20000 years ago, still in the Russian steppe. Although the didn’t begin to migrate westwards until much later, the desc_LINES 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.

WARNING!
The details of this map cannot currently be proved with any scientific rigour. The exact path is not meant to be a true representation of historical migrations: it adopts the minimum path length that adequately describes the current data. Details of some alternative hypotheses are given in the descriptive text, and numerous other, more authoritative charts are available elsewhere.
AGE ESTIMATION FROM NEXT-GEN TESTS (BIG Y, ETC.)

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 “next-generation” tests like BigY or YElite 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 taking the number of SNP mutations that are not shared by all the members of that clade (m) and dividing it by the timescale for SNP formation (r) and multiplying it by the number of testers (n), thus:

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

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

ACCOUNTING FOR SMALL-NUMBER STATISTICS

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

For large clades, calculating this uncertainty becomes technically impractical, so we use the Gaussian approximation that the uncertainty in m is the square root of m, and that the uncertainty in \( \frac{1}{m} \) is 1.96 x sqrt(\( \frac{1}{m} \)).

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:

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

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 a mutation like U106 can be 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 \( n \) testers which have been tested. This is the “time to most-recent common ancestor” or TMRCA. This is subtly different from the SNP age: the actual age of the quoted SNP. In most cases, this distinction doesn’t matter, but it can become important in some clades.

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

In this case, A, B and C share a most-recent common ancestor (ABC) and a terminal common SNP (X). The age of X is slightly older than that 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:

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

- **B:** X+ Y+ Z+ 1 novel variant
- **C:** X+ Y+ Z+ 1 novel variant

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

If tester A then comes along with the results:

- **A:** X+ Y+ Z− 2 novel variants

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

Sometimes additional data is available (e.g. from a different next-generation test 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.

At the present time (Jan 2016), we are only calculating dates from BigY tests, as we have both a sufficient number of these, and a good enough idea of their calibration to do so. The calibration exists to do the same analysis with the FGC YElite tests, but this remains a future exercise.

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 process described above. However, they will occur at a fixed time before the TMRCA or convergence age. This is given by:

\[ t = \frac{1}{r} \left( \frac{n}{2} - \frac{1}{2} \right) \]

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 \( \pm 0.475 \times \sqrt{n} \) for large \( n \).
FINAL AGE CALCULATION

The final age is determined from three numbers:

Firstly, from the number of SNPs beneath the target:

\[ t = m/r \]  \[ \text{[T1]} \]

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

\[ t_0 = n(U106) - m_0r_0 \]  \[ \text{[T2]} \]

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

\[ r_0 = m(U106)/r(U106) \]  \[ \text{[R2]} \]

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

\[ t_p = t_0(p) - m_p r_0 \]  \[ \text{[T3]} \]

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

\[ t = m r_0/n \]  \[ \text{[T4]} \]

which then gives the equality:

\[ t_0 = t_p \]  \[ \text{[T5]} \]

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 (see "Defining the Present Day", below).

FINAL AGE UNCERTAINTY

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

\[ \delta t = 1.96 \sqrt{\left( \frac{\delta n/m}{m} \right)^2 + \left( \frac{\delta r}{r} \right)^2} \]  \[ \text{[ET1]} \]

where \( \delta t \) is derived from the literature or case studies. Similiarly, the uncertainty in \[ \text{[T4]} \] can be derived as:

\[ \delta t = 1.96 \sqrt{\left( \frac{\delta n/m}{m} \right)^2 + \left( \frac{\delta r_0}{1} \right)^2} \]  \[ \text{[ET2]} \]

where \( \delta r_0 \) is given from \[ \text{[R2]} \] by:

\[ \delta r_0 = 1.96 (m(U106)/r(U106)) \]  \[ \text{[ER1]} \]

In both cases, \( m-n \delta m \) and \( m+n \delta m \) are given by the highest and lowest value of \( \lambda \), respectively, for which:

\[ I_{\lambda-\alpha}^{\alpha} \text{Pois}(\lambda,k) = 0.1585 \]  \[ \text{[ER2]} \]

\[ I_{\lambda-\alpha}^{\alpha} \text{Pois}(\lambda,k) = 0.8415 \]  \[ \text{[ER3]} \]

at 1σ and:

\[ I_{\lambda-\alpha}^{\alpha} \text{Pois}(\lambda,k) = 0.025 \]  \[ \text{[ER4]} \]

\[ I_{\lambda-\alpha}^{\alpha} \text{Pois}(\lambda,k) = 0.975 \]  \[ \text{[ER5]} \]

at 95% confidence, where \( \text{Pois}() \) is the Poisson function, \( \lambda/\sqrt{\lambda} \). For large \( m \), where this value is computationally expensive to determine, the approximation \( \delta m = 1.96 /m \) 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 \( n \) and is calculated in time since \( \pm 1.96 \delta \) for U106 and the parent SNP, respectively, rather than from the present day.

If \( \delta k < \delta k \) (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 is based on the parent SNP. This provides an age propagated forward in time, which we will call \( t_\lambda \) (uncertainty \( \delta t_\lambda \)). Note that as \[ \text{[ER4]} \] and \[ \text{[ER5]} \] provide asymmetric errors around \( m \), the final uncertainty, \( \delta t_\lambda \), will be asymmetric around \( t_\lambda \) as well.

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

\[ \delta t_{\text{final}} = \left( \frac{t_\lambda + \delta t_\lambda}{w} \right) \]  \[ \text{[ET3]} \]

where the weights are set as follows:

\[ w = \frac{(t\lambda/\lambda)^2}{(t\lambda/\lambda)^2 + (\delta t_\lambda/\lambda)^2} \]  \[ \text{[ET4]} \]

\[ w_2 = \frac{(t\lambda/\lambda)^2}{(t\lambda/\lambda)^2 + (\delta t_\lambda/\lambda)^2} \]  \[ \text{[ET5]} \]

where \( t\lambda \) is either \( t_\lambda \) or \( n(U106) \), depending on which gives the more accurate age. The same limits are applied such that the cluster cannot be older than its parent and cannot be younger than the present day.

DEFINING THE PRESENT DAY

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

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

CHOOSING A MUTATION RATE

We have so far ignored how the choice of the underlying mutation rate, \( r \), and its uncertainty, \( \delta r \), are calculated. Ultimately, these come from three sources: (1) counting SNPs from known lineages among the BigY tests themselves, (2) literature studies which perform the same task of summing up a measured number of mutations which have occurred over a known period of time, (3) limits from DNA testing of archaeological remains.

In the next section, a selection of these are applied to the BigY tests. 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.

(1) RATES FROM LINEAGES IN BIGY

As of Jan 2016, we have 103 BigY tests from lineages where we have a named individual who is very likely the common ancestor of at least two tests, where we also have positions of the mutations accumulated since that common ancestor. “Very likely” in this case is a judgement call made based on paper-trail and genetic evidence. In total, they represent around 46,000 years of lineages.

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

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

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

BigY tests from other families only account for around 10,000 years of lineages. Although the uncertainties on these are larger, they roughly show the same rate as the bulk literature, as illustrated on the next page.
RATES FROM THE LITERATURE

A growing number of studies are performing thorough analyses of the human genome mutation rate. Only a few of these directly provide rates specific to the Y chromosome. 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. Such studies include Xue et al. (2009) and Helgason et al. (2015). Helgason et al. uniquely provide two estimates, for the palindromic and non-palindromic regions, which show a marginal difference in mutation rate of around 18%. The slower palindromic rate is consistent with the paternally transmitted autosomal rate.

Some studies measure the rate of autosomal mutations in father–mother–child (“triplet”) groups and scale this rate to the Y chromosome, based on the ratio of expected mutations inherited from the father and the mother. Examples include Mendez et al. (2013) and Scozzari et al. (2013). These rates are consistent with the slower (palindromic) rate from Helgason et al.

A third method involves taking an genetic-cultural group with a known date of origin, and computing a mutation rate based on that age. Examples include Poznik et al. (2013) for Native Americans, and Francalacci et al. (2013) for Sardinians. Here, we add an extra 10–20% uncertainty to their rates, to reflect the uncertainty in the date at which the tested population formed.

A final method relies on ancient DNA from archaeological remains. Such remains include the Ust'-Ishim burial in western Siberia, Malta’s Boy and Kostenki 14. Together, these represent over 100,000 years of lineage. Examples include Fu et al. (2014), Karmin et al. (2015) and Trombetta et al. (2015). As these results rely on the archaeological remains themselves, care should be taken when consistency among these different methods clearly shows that the mutation rate has not changed significantly over tens of millennia.

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.

Typically, ancient DNA results will return an age based on 14C dating, and a haplogroup. Where a study using a known rate has counted the number of SNPs in a modern population of the same haplogroup (e.g. Hallast et al. 2014), a new rate can be estimated based on the ratio of their age to the 14C age, multiplied by the mutation rate they assume. (Note that 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.) Normally only limits can be found from archaeological remains, except in cases where novel variants can be accurately found and counted in the ancient DNA.

COMBINING RATES FROM DIFFERENT ESTIMATES

We can use weighted averages to combine the estimates provided by different studies, while taking care to avoid circular reasoning. The chart below shows the mutation rate in SNPs per year per base pair. It can be read as follows:

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Estimate</th>
<th>Date Range</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BigY</strong></td>
<td>0.3</td>
<td>(3000–5000)</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>Literature estimates</strong></td>
<td>0.4</td>
<td>(10000–1000)</td>
<td>0.6</td>
</tr>
<tr>
<td><strong>Poznik et al. (2013)</strong></td>
<td>0.5</td>
<td>(1000–5000)</td>
<td>0.4</td>
</tr>
<tr>
<td><strong>Helgason et al. (2015)</strong></td>
<td>0.6</td>
<td>(1000–5000)</td>
<td>0.3</td>
</tr>
<tr>
<td><strong>Mendez et al. (2013)</strong></td>
<td>0.7</td>
<td>(1000–5000)</td>
<td>0.2</td>
</tr>
<tr>
<td><strong>Scozzari et al. (2013)</strong></td>
<td>0.8</td>
<td>(1000–5000)</td>
<td>0.1</td>
</tr>
</tbody>
</table>

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.

This result in a rate which (as of 21 Jan 2016) is 198 years/SNP, with a 95% confidence interval of 181–209 years/SNP, for the euchromatic, non-palindromic region of BigY tests. The full test minus DYZ19 is 129 years/SNP (125–139 years/SNP) and for all bases reported in the BigY test is 125 years/SNP (116–137 years/SNP).

COMPARISON TO YFULL

YFull.com also operate their own age-dating system, which works on a similar basis to ours. The major differences are that we apply more rigorous causality checks, particularly to our uncertainty estimates and, conversely, YFull has the luxury of making its own variant identification from the BAM file. YFull assumes a coverage of 8467165 base pairs compared to our 8753254. Our ages compare to theirs as follows (ages obtained 21 Jan 2016):

<table>
<thead>
<tr>
<th>YFull (years)</th>
<th>Our data (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M269</td>
<td>6400 (7300–5500)</td>
</tr>
<tr>
<td>L13</td>
<td>6200 (6900–5600)</td>
</tr>
<tr>
<td>L151</td>
<td>5800 (6400–5200)</td>
</tr>
<tr>
<td>L151</td>
<td>4900 (5400–4500)</td>
</tr>
<tr>
<td>U106</td>
<td>4900 (5400–4500)</td>
</tr>
<tr>
<td>Z18</td>
<td>3500 (4000–3100)</td>
</tr>
<tr>
<td>Z732</td>
<td>3500 (3800–2200)</td>
</tr>
<tr>
<td>Z156</td>
<td>2800 (4300–1700)</td>
</tr>
<tr>
<td>Z381</td>
<td>4900 (5400–4500)</td>
</tr>
<tr>
<td>Z304</td>
<td>4200 (4700–3600)</td>
</tr>
<tr>
<td>Z301</td>
<td>4800 (5400–4200)</td>
</tr>
<tr>
<td>L48</td>
<td>4800 (5400–4200)</td>
</tr>
<tr>
<td>L47</td>
<td>4500 (3300–3800)</td>
</tr>
<tr>
<td>Z9</td>
<td>4800 (5400–4200)</td>
</tr>
<tr>
<td>Z22</td>
<td>4200 (4800–3700)</td>
</tr>
<tr>
<td>Z8</td>
<td>2900 (3500–2300)</td>
</tr>
</tbody>
</table>

Notably almost all our ages agree within the uncertainties. The only exception, Z9, is very close and statistically acceptable in this ensemble.

Which is the more accurate test is a complex question to answer: we have more tests, but YFull can perform better and more homogeneous analysis from the BAM file directly. For well-populated clades (>20 tests in YFull), where the full known branching structure is present in the YFull tree, the YFull ages should be more accurate. For small clades, or where branches not present in the YFull tree, our ages should be more accurate. The absolute calibration of both systems matches very well at the U106 level.
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 negligible, 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 will use 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 will we discuss later.

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

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

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

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

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

**INTRA-CLADE AND INTER-CLADE AGES**

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

Intra-clade ages are generally problematic, as they ignore the fact that many people within a clade are closely related: e.g., many calculations of the intra-clade age of U106 will be biased by the fact that half of people are L48+.

The calculated age will be pulled down towards the L48 age. For this reason, inter-clade ages are generally used for STR calculations, which compare two clades to each other.

**FINITE AGE COMBINATIONS**

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

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

The final age is therefore given by:

\[
I_{AB} = \frac{I_{AA} + I_{BB} + \ldots + I_{AB} + I_{BA}}{n} \]

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

\[
\sigma^2 = \frac{\sigma(AB)^2}{mn} + \left(\frac{\sigma(\mu)^2}{\sigma_{\mu}^2}\right)^2 \]

where \(\sigma(AB)\) is the standard deviation among all TMRCAs in the A–B set, \(\mu\) is the mutation rate on marker \(i\) and \(\sigma(\mu)|_{\mu = \mu_i}\) is a weighting factor which is the fraction of test pairs which are compared on marker \(i\). The left-hand ratio represents the square of the standard error in the mean, and the right-hand ratio represents the square of the fractional uncertainty in the mutation rate. The square root of this gives \(\sigma_{IA}\), the uncertainty in \(I_{AA}\).

**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:

\[
\text{Var}(AB) = \frac{s_{A}^2}{m} + \frac{s_{B}^2}{n} - \frac{s_{AB}}{mn} \]

where \(s_{A} = \sigma(\mu)^2/x_{A}\), and similarly for \(s_{B}\) and \(s_{AB}\) for \(j = 1\) to \(n\). The square of the fractional uncertainty in that variance (at the 68% confidence interval) will be:

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

Variances on individual markers can be summed, such that:

\[
\text{Var}(AB) = \sum_{i} \text{Var}(AB) \]

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

\[
\sigma(\text{Var}(AB))/\text{Var}(AB) = \left(\frac{\sigma(\text{Var}(AB))}{\text{Var}(AB)}\right) / 111
\]

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

\[
\tau(AB) = \text{Var}(AB) \sum_{i} (\sigma(\mu))^2 / 2\]

and:

\[
\sigma(\tau(AB)) = \tau(AB) \left[\frac{\sigma(\text{Var}(AB)/\text{Var}(AB))^2}{\left(\sum_{i} \sigma(\mu))^2}\right] \right]
\]

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

**CHOICE OF MARKERS**

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

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

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of Markers</th>
<th>Number 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.

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

CONSTRAINTS APPLIED IN THE FOLLOWING

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

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

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

DATA USED IN THE ANALYSIS

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

CALIBRATION OF STR TO SNP AGES: VARIANCES

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

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

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

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

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

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

CALIBRATION OF STR TO SNP AGES: INFINITE ALLELES

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

The following fitting parameters are derived:

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

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.
**DESCRIPTION**

This phylogenetic tree of U106 shows the relationships between the 602 U106 and 1 U106c family members, with a focus on the SNP, U106. The tree illustrates the evolution of the SNP across different populations and time periods, highlighting key events and migrations. Each node represents a specific SNP mutation, and the branches indicate the evolutionary relationships between these mutations.

**Archeological M209 remains in modern Russia**
- *Peres Oscillation* (Kurgan wave 3: steppe → E. Europe)
- *Corded wave in C. Europe*

**Portugal: First Bell Beaker culture**
- Wave 3: steppe → E. Europe
- Corded wave in C. Europe

**U106 family tree**

**Update:** 10 Mar 2016, Dr. Jon McDougall for the U106/S2 group

**CONVERGENCE DATES**

- 2200 BC
- 2100 BC
- 2000 BC
- 1900 BC
- 1800 BC
- 1700 BC
- 1600 BC
- 1500 BC
- 1400 BC
- 1300 BC
- 1200 BC
- 1100 BC
- 1000 BC

**Estonia: Kaali impact?**
- 1100 AD

**Russia: Viking invasions**
- 1000 AD

**Iberia: Reconquista**
- 1100 AD

**England: Aylesbury**
- 1000 AD

**Spain: Battle of Tábara**
- 1000 AD

**Catalan War**
- 1000 AD

**Iberian: Golden age**
- 900 AD

**Start of Iron Age in C. Europe**
- 800 AD

**Mediterranean: Early Colonisation**
- 800 AD

**Estonia: Raapina culture**
- 700 AD

**England: Aylesbury**
- 600 AD

**England: Edward I's Crusade**
- 600 AD

**Hebrew Exodus**
- 600 AD

**Late Bronze Age collapse**
- 600 AD

**Bronze Age reaches Ireland**
- 500 AD

**England: and of Devizes-Ridgeway war**
- 500 AD

**Scotland: Trzciniec culture**
- 400 AD

**England: Aylesbury**
- 300 AD

**France: Battle of Poitiers**
- 200 AD

**England: Aylesbury**
- 100 BC

**Portugal: First Bell Beaker culture**
- 900 BC

**Corded wave in C. Europe**
- 800 BC

**Mediterranean: Early Colonisation**
- 700 BC

**Britain: Early Colonisation**
- 600 BC

**England: Aylesbury**
- 500 BC

**England: Aylesbury**
- 400 BC

**England: Aylesbury**
- 300 BC

**England: Aylesbury**
- 200 BC

**England: Aylesbury**
- 100 BC

**England: Aylesbury**
- 1 AD

**England: Aylesbury**
- 10 AD

**England: Aylesbury**
- 20 AD

**England: Aylesbury**
- 30 AD

**England: Aylesbury**
- 40 AD

**England: Aylesbury**
- 50 AD

**England: Aylesbury**
- 60 AD

**England: Aylesbury**
- 70 AD

**England: Aylesbury**
- 80 AD

**England: Aylesbury**
- 90 AD

**England: Aylesbury**
- 100 AD

**England: Aylesbury**
- 110 AD

**England: Aylesbury**
- 120 AD

**England: Aylesbury**
- 130 AD

**England: Aylesbury**
- 140 AD

**England: Aylesbury**
- 150 AD

**England: Aylesbury**
- 160 AD

**England: Aylesbury**
- 170 AD

**England: Aylesbury**
- 180 AD

**England: Aylesbury**
- 190 AD

**England: Aylesbury**
- 200 AD

**England: Aylesbury**
- 210 AD

**England: Aylesbury**
- 220 AD

**England: Aylesbury**
- 230 AD

**England: Aylesbury**
- 240 AD

**England: Aylesbury**
- 250 AD

**England: Aylesbury**
- 260 AD

**England: Aylesbury**
- 270 AD

**England: Aylesbury**
- 280 AD

**England: Aylesbury**
- 290 AD

**England: Aylesbury**
- 300 AD

**England: Aylesbury**
- 310 AD

**England: Aylesbury**
- 320 AD

**England: Aylesbury**
- 330 AD

**England: Aylesbury**
- 340 AD

**England: Aylesbury**
- 350 AD

**England: Aylesbury**
- 360 AD

**England: Aylesbury**
- 370 AD

**England: Aylesbury**
- 380 AD

**England: Aylesbury**
- 390 AD

**England: Aylesbury**
- 400 AD

**England: Aylesbury**
- 410 AD

**England: Aylesbury**
- 420 AD

**England: Aylesbury**
- 430 AD

**England: Aylesbury**
- 440 AD

**England: Aylesbury**
- 450 AD

**England: Aylesbury**
- 460 AD

**England: Aylesbury**
- 470 AD

**England: Aylesbury**
- 480 AD

**England: Aylesbury**
- 490 AD

**England: Aylesbury**
- 500 AD

**England: Aylesbury**
- 510 AD

**England: Aylesbury**
- 520 AD

**England: Aylesbury**
- 530 AD

**England: Aylesbury**
- 540 AD

**England: Aylesbury**
- 550 AD

**England: Aylesbury**
- 560 AD

**England: Aylesbury**
- 570 AD

**England: Aylesbury**
- 580 AD

**England: Aylesbury**
- 590 AD

**England: Aylesbury**
- 600 AD

**England: Aylesbury**
- 610 AD

**England: Aylesbury**
- 620 AD

**England: Aylesbury**
- 630 AD

**England: Aylesbury**
- 640 AD

**England: Aylesbury**
- 650 AD

**England: Aylesbury**
- 660 AD

**England: Aylesbury**
- 670 AD

**England: Aylesbury**
- 680 AD

**England: Aylesbury**
- 690 AD

**England: Aylesbury**
- 700 AD

**England: Aylesbury**
- 710 AD

**England: Aylesbury**
- 720 AD

**England: Aylesbury**
- 730 AD

**England: Aylesbury**
- 740 AD

**England: Aylesbury**
- 750 AD

**England: Aylesbury**
- 760 AD

**England: Aylesbury**
- 770 AD

**England: Aylesbury**
- 780 AD

**England: Aylesbury**
- 790 AD

**England: Aylesbury**
- 800 AD

**England: Aylesbury**
- 810 AD

**England: Aylesbury**
- 820 AD

**England: Aylesbury**
- 830 AD

**England: Aylesbury**
- 840 AD

**England: Aylesbury**
- 850 AD

**England: Aylesbury**
- 860 AD

**England: Aylesbury**
- 870 AD

**England: Aylesbury**
- 880 AD

**England: Aylesbury**
- 890 AD

**England: Aylesbury**
- 900 AD

**England: Aylesbury**
- 910 AD

**England: Aylesbury**
- 920 AD

**England: Aylesbury**
- 930 AD

**England: Aylesbury**
- 940 AD

**England: Aylesbury**
- 950 AD

**England: Aylesbury**
- 960 AD

**England: Aylesbury**
- 970 AD

**England: Aylesbury**
- 980 AD

**England: Aylesbury**
- 990 AD

**England: Aylesbury**
- 1000 AD

**England: Aylesbury**
- 1010 AD

**England: Aylesbury**
- 1020 AD

**England: Aylesbury**
- 1030 AD

**England: Aylesbury**
- 1040 AD

**England: Aylesbury**
- 1050 AD

**England: Aylesbury**
- 1060 AD

**England: Aylesbury**
- 1070 AD

**England: Aylesbury**
- 1080 AD

**England: Aylesbury**
- 1090 AD

**England: Aylesbury**
- 1100 AD

**England: Aylesbury**
- 1110 AD

**England: Aylesbury**
- 1120 AD

**England: Aylesbury**
- 1130 AD

**England: Aylesbury**
- 1140 AD
Geographical distribution of U106

Updated: 11 February 2016
Dr. Iain McDonald
on behalf of the U106/S21 group

Distribution of all U106 by region and sub-clade

This represents the geographical and phylogenetic distribution of 1613 U106+ tests from Family Tree DNA. These include members from the U106, U198 and L1 haplogroup projects, and several geographical projects. Information was collected during February 2016. Geography is self-reported by the testers. Phylogenetic position is based on the last SNP tested, thus some testers may branch into lower sub-clades for which they have not been tested.

Overall European structure

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.

England 543 testers
Scotland 165 testers
Wales 17 testers
Ireland 156 testers
Norway 58 testers
Sweden 77 testers
Finland 25 testers
BeNeLax 98 testers
Germany 190 testers
France 98 testers
Denmark 24 testers
Iberia 11 testers
Switzerland 23 testers
Italy 23 testers
South-East Europe 17 testers

Former USSR 28 testers
Excludes Baltic States but includes Russia, Belarus & the Ukraine.

Baltic States 60 testers
Covers Poland, Estonia, Latvia, Lithuania and Russian Kaliningrad

East-Central Europe 39 testers
Covers Austria, Czech Republic, Slovakia & Hungary

Baltic States

Former USSR

Distribution of all U106 by region and sub-clade

This represents the geographical and phylogenetic distribution of 1613 U106+ tests from Family Tree DNA. These include members from the U106, U198 and L1 haplogroup projects, and several geographical projects. Information was collected during February 2016. Geography is self-reported by the testers. Phylogenetic position is based on the last SNP tested, thus some testers may branch into lower sub-clades for which they have not been tested.

○ = 1 tested person

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

Overall European structure

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.

Distribution of all U106 by region and sub-clade

This represents the geographical and phylogenetic distribution of 1613 U106+ tests from Family Tree DNA. These include members from the U106, U198 and L1 haplogroup projects, and several geographical projects. Information was collected during February 2016. Geography is self-reported by the testers. Phylogenetic position is based on the last SNP tested, thus some testers may branch into lower sub-clades for which they have not been tested.

○ = 1 tested person

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

Overall European structure

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.

Distribution of all U106 by region and sub-clade

This represents the geographical and phylogenetic distribution of 1613 U106+ tests from Family Tree DNA. These include members from the U106, U198 and L1 haplogroup projects, and several geographical projects. Information was collected during February 2016. Geography is self-reported by the testers. Phylogenetic position is based on the last SNP tested, thus some testers may branch into lower sub-clades for which they have not been tested.

○ = 1 tested person

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

Overall European structure

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.

Distribution of all U106 by region and sub-clade

This represents the geographical and phylogenetic distribution of 1613 U106+ tests from Family Tree DNA. These include members from the U106, U198 and L1 haplogroup projects, and several geographical projects. Information was collected during February 2016. Geography is self-reported by the testers. Phylogenetic position is based on the last SNP tested, thus some testers may branch into lower sub-clades for which they have not been tested.

○ = 1 tested person

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

Overall European structure

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.

Distribution of all U106 by region and sub-clade

This represents the geographical and phylogenetic distribution of 1613 U106+ tests from Family Tree DNA. These include members from the U106, U198 and L1 haplogroup projects, and several geographical projects. Information was collected during February 2016. Geography is self-reported by the testers. Phylogenetic position is based on the last SNP tested, thus some testers may branch into lower sub-clades for which they have not been tested.

○ = 1 tested person

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

Overall European structure

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.

Distribution of all U106 by region and sub-clade

This represents the geographical and phylogenetic distribution of 1613 U106+ tests from Family Tree DNA. These include members from the U106, U198 and L1 haplogroup projects, and several geographical projects. Information was collected during February 2016. Geography is self-reported by the testers. Phylogenetic position is based on the last SNP tested, thus some testers may branch into lower sub-clades for which they have not been tested.

○ = 1 tested person

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

Overall European structure

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

METHODS TO IDENTIFY MIGRATION PATHS

There are three primary ways to work out how and when a clade spread.

1. Look at the current distribution of people. This tells you something about who is related to who, but not when and why.

2. Look at archaeological DNA results. As already discussed, this is very good at determining upper limits to when clades formed. It can also tell you something about the earliest phases of a population’s presence in an area. However, these are only glimpses: snapshots into a forgotten world, and very few and far between. There’s only so much information they can give on particular time periods and particular migrations, unless they happen to be very large. Nevertheless, this is the most effective method for older populations.

3. Look at the ages of MRCAs from different countries. This is perhaps the most powerful tool for more recent populations, but perhaps also the most difficult to obtain good data from. We will discuss this later.

CURRENT DISTRIBUTION

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

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

ARCHAEOLOGICAL Y-DNA

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

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

THE ROUTE INTO EUROPE

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

LOCATION OF THE FIRST U106

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

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

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

Later pages in this section show this formation and dispersion of U106 graphically.
6500-4000 BC: the situation in Europe

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.

**Loschbour**
- **I2a1**
- Haak+ 2015
- 6220-5990 BC

**La Branya**
- **C1a2**
- Olande 2014
- 5940-5690 BC

**Motala**
- **I2c2, various I2a1**
- Haak+ 2015
- 5898-5531 BC

**Sok River**
- **R1b1-M343, L278**
- (xM478, M269)
- Haak+ 2015
- 5650-5555 BC

**Yuzhnyy Oleni Ostrov**
- **R1a1-M459, Pages65.2**
- Haak+ 2015
- 5500-5000 BC

**Serteya**
- **R1a1**
- Chekunova 2014
- 4000 BC

**Alsónyék-Bátaszék & Lánycsók**
- **2*F (GHIJK?), H2, G2, 3*G2a, G2a2b, I**
- Szécsényi-Nagy 2014
- various dates between 5840-5550 BC

**Tiszaszölös-Domaha´za**
- **I2a**
- Szécsényi-Nagy 2014
- 5780-5650 BC

**Vinkovci & Vukovar**
- **G2a, G2a, I2a1**
- Szécsényi-Nagy 2014
- est. 6000-5500 BC

**Els Troncs**
- **F (xG, …), I2a1b1**
- R1b1-M343(xM269)
- Haak 2015
- 5311-5068 BC

**Tolna-Mözs**
- **G2a2b, F-M89 (GHIJK?)**
- Szécsényi-Nagy 2014
- various dates between 5300-5020 BC

**Berekalja**
- **C1a2**
- Szécsenyi-Nagy 2014
- 5300-4950 BC

**Derenburg**
- **2*F-M89 (xGHIJK), G2a2b**
- Brandt 2013
- 5300 - c. 5000 BC

**LBK**
- **Halberstadt**
- 3*G2a2a; 2*G2a2a1
- Haak+ 2015
- various: 5207-4946 BC

**Kompolt-Kigyoser**
- **C1a2**
- Gamba 2014
- 5210-4990 BC

**Late ALP**
- **Szőlőskert­Tangazdaság**
- **G2a2b**
- Szécsényi-Nagy 2014
- est. 5100 BC

**LBKT**
- **Karsdorf**
- **T1a**
- Haak+ 2015
- 5207-5070 BC

**Balatonszemes-Bagódomb**
- **I1-M253**
- Szécsényi-Nagy 2014
- est. 5000 BC

**LBKT**
- **Avellaner**
- **3*G2a, E1b-M35.1+ V13+**
- Lacan 2011b
- 5000 BC

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

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

11300 BC (12900 BC - 9700 BC)
DESCRIPTION
This page details the archaeological DNA obtained from burials between 4000 and 3000 BC, which show Europe during the initial R-M269 expansion, before U106 formed. Below are the symbols used in this page:

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

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

### Western Europe:
- **Megalithic Quedlinburg**: 3645-3537 BC, R-P224?
- **Trielles**: 3000 BC, 2*G2a-P15, 2*I2a1-M438,P37.2
- **Remedello di Sotto**: I2 [I2a1a], 3483-3107 BC
- **Yamnaya Ishkinovka**: R1b-M269>L23(xZ2015), 3300-2700 BC
- **Yamnaya Ishkinovka**: R1b-M269>L23(Z2015), 3300-2700 BC
- **Kutuluk**: R1b-M269>L23(Z2015), 3300-2700 BC
- **Lopatino**: R1b1a2-M269(xL51), 3339-2917 BC
- **Yamnaya Peshany**: R1b1a2-M269(xL51), 3334-2635 BC

### Central Europe:
- **Serteya**: R1a1, 4000 BC
- **La Mina**: 2*I2a1a1 (or 1 H2?), 3900-3600 BC
- **Esperstedt**: 3887 - 3797 BC, I2a1b1a
- **Esperstedt**: 3360 - 3086 BC, I2a1b1a
- **Salzmünde/Shiebzig**: 3400 - 3025 BC, G2a2a
- **F-P316* (xGHIJLNOP)**, 3887 - 3797 BC
- **Baalburg, Salzmünde/Bernburg**: TRB, 3645-3537 BC

### Russia:
- **Russia: 100% R**
- **Lopatino**: R1b1a2-M269(xL51), 3339-2917 BC
- **Yamnaya Ishkinovka**: R1b-M269>L23(Z2015), 3300-2700 BC
- **Kutuluk**: R1b-M269>L23(Z2015), 3300-2700 BC

### Other haplogroups:
- **R1a**: 81% G, 19% I
- **R1b**: 50% I, 33% G, 17% R?
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.

**3000-2500 BC: M269 invades Europe**

- **Southern Scandinavia:**
  - 67% R
  - 33% I

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

- **Russia:**
  - 83% R
  - 8% N
  - 8% I

- **Southern & Western Europe:**
  - 77% G
  - 23% I

- **El Mirador:**
  - 1*I2a1, 3*I2a2, 1*I, 2*G2a

- **Kromsdorf:**
  - R1b-M269xU106, R1b-M343(M269?)

- **Ulan:**
  - I2a [I2a2a1b1b2-S12195]

- **Kyndelose:**
  - R1a [R1a1a1]

- **Bergheinfeld:**
  - R1a [R1a1a1-M417xZ647]

- **Viby:**
  - R1a [R1a1a1]

- **Oblaczkowo:**
  - R1b1

- **Eulau:**
  - R1a1

- **Stalingrad Quarry:**
  - R1b [Z2107~Z2105]

- **Jagodno:**
  - G?, I or J?

- **Jagodno:**
  - G?, I or J?

- **Kutuluk:**
  - R1b-M269>L23>Z2015

- **Dolmen de La Pierre Fritte:**
  - 2*I2a1

- **Peshany:**
  - R1b1a2-M269

- **Remedello di Sotto:**
  - 2*I2 [I2a1a1a-L672/S327]

- **Ajvide:**
  - 2800-2000 BC

- **Naumovo:**
  - R1a1

- **Zhizhitskaya:**
  - R1a1 + N1c

- **Termta:**
  - 3*R1b [PF6482~M269, CTS9416~Z2105, Z2105]
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.

2500-2000 BC: early growth of U106

Southern Scandinavia:

- 75% R
- 25% I

Russia:

- 67% R
- 16% N
- 16% I

Central Europe:

- 54% R
- 46% I

Update: 17 June 2015
Dr. Iain McDonald on behalf of the U106/S21 group
The origins of U106: 4400 BC to 2900 BC

(1) Arrival into Europe
The origin of U106 may now be placed around 2900 BC. There is roughly a 2-in-3 chance of it being within 300 years of this date. Ancient DNA shows few haplogroup R men in Europe before about 3000 BC, so it is known from branches/further up the tree and archaeological results that haplogroup R was not in Asia. We can presume that U106 was founded somewhere late in this migration from Asia to Europe.

(2) Kurgan hypothesis
Key mutations often arise during population expansion events. These will typically shortly precede (or occur during) migration events when one group takes over another. The important Asia–Europe migration took a migration from the east to the west. The Kurgan expansion is the only known, major migration that fits both the likely range of dates and the east–west movement. In addition it appears to fit the trans-Ural area near concentrations of groups further up the haplogroup R1b tree (e.g. V88).

(3) Upstream SNPs
The geographical median of SNPs between M269 and U106 follows a east-west trend in surveys of both modern and ancient DNA. We can use this to infer that the M269→U106 sequence follows an ~800-year gap in our knowledge. This may represent a hiatus in the east–west population movement. In addition it appears to arise from the trans-Ural region shown here, or perhaps even further east.

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

(5) M269
M269 formed around 4400 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 Donets, took place around 4500–4000 BC and could be an origin for M269. Wave 2 probably occurred from the Maykop culture (1700–3000 BC, indicated in cyan). The high variance and unusual M269 population found in this area (Hovhannisyan et al. 2004).

(6) M269→L23–L51
Hovhannisyan et al. (2014, fig. 8A) notes a discontinuity between the Neolithic, Bactrian, Turkic–Armenian and European M269 populations. L51 does not clearly appear in the ancient DNA results, although ancient DNA from the Yamnaya culture (modern day Russia and Ukraine) shows that L23 and its subclade Z2013 dominate here. Additionally, FTDNA shows M269→L23 and L23→L51 now are much more focussed on the Black Sea than L51+ testers. We interpret this as indicating L51 formed after a migration started heading from the Black Sea towards the European peninsula.

7) M269→L23–L51
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 subsequent westward expansion. PT7500 and its subclades are clearly concentrated in the Turkish highlands, CTS8782, particularly CTS9219, is closely associated with the Balkan peninsula, L272 is spread around eastern Europe; while CTS7763 may be Balkan or Anatolian. Meanwhile, ancient DNA from Holo et al. (2013) shows a very strong replacement of Z2013 throughout modern-day Russia. These data suggest Z2013 went north, south and east, while L51 went west.

(8) Caucaian populations
There are few cases where specific SNPs have localised among people from the Caucasus, at least at FTDNA. Two M269 testers are L23→Z2013 and L23→L51→P313. Combined with the Hovhannisyan and ancient DNA (Holo et al. results), we suggest they are probably mostly L23→Z2013.

(9) M269→L23–L51→PT7500 and the Hittites
The variance of PT7500 in Turkey shares the European focus of L51→P311 but is not widely found in Germany. The median location of PT7500 in FTDNA tests is close to the Black Sea and so barely south of the project. The east–west migration means that the split between P311 and PT7500 is likely to be several thousand years old, although L51 may represent the population that matched into Europe; they perhaps (initially) did not get very far.

(10) L51→PT7500 migration into Europe
L51 through to P311 represents an ~800-year gap in our knowledge. This may represent a hiatus in the east–west population movement. In addition it appears to fit the trans-Ural area near concentrations of groups further up the haplogroup R1b tree (e.g. V88).

(11) The path into Europe: North or South?
The path our ancestors took into Europe isn't well defined. Traveled in prehistoric Europe was presumably by watercourses, driven by mountain ranges, dense forestation and marshland. Two main options are considered. Firstly, a path through the Danube river, which leads directly from the Black Sea to Germany, while P311 is first found in the archaeological and where there is the east–west most substantial concentration of P311 in the present population. This better fits the distribution of M269→P311 clades, which is clearly centered in the east–west trend in surveys of both modern and ancient DNA. We can use this to infer that the M269→U106 sequence follows an ~800-year gap in our knowledge. This may represent a hiatus in the east–west population movement. In addition it appears to arise from the trans-Ural region shown here, or perhaps even further east.

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

(13) P312 and U106
P312 is U106's bigger (though not necessarily older) brother. It's worth considering how P312 fared after the P312–U106 migration. In addition it appears to arise from the trans-Ural region shown here, or perhaps even 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, and Denmark for the northern route. We have placed this foundation around 2900 BC, give or take a few centuries, which allows U106 to form part of the Single Grave, Battle Axe and Bell Beaker cultures, implying that the pre- and proto-Celtic cultures that followed them were similar in U106.

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

(15) The path into Europe remains ill defined. On balance, my personal consideration is that the northern route has the better evidence.
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, whereas twins sire 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's will be a generation younger. Of each of those halves, half will be related another generation down, etc. So we end up with a histogram that halves with each generation, like the one above.

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

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

These effects are subtle, and best illustrated through a real-world example. Here, we consider two examples, U106>Z381>Z301>L48>Z9>Z30>Z2>Z7>Z8 and U106>Z381>Z156>DF98.

These effects are subtle, and best illustrated through a real-world example. Here, we consider two examples, U106>Z381>Z301>L48>Z9>Z30>Z2>Z7>Z8 and U106>Z381>Z156>DF98.

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 TMRCA's for DF98–Z8 pairs*, we arrive at the following histogram (binned into 20-year bins):

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

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

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

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

In these inter-clade 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, going 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 histograms, we can form an expectation of what one should look like.

In the simplified population presented previously, we have a slowly growing population, with a new branch forming every 90-140 years or so, depending on where 67 or 111 markers are being tested (let’s say 120 years). Half of the population goes into each branch, so half of the TMRCAs 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.

Intra-clade examples

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

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

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

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

Comparing geographical regions

The major advantage of this test comes when we start to split up these intra-clade TMRCAs by region to provide inter-regional and intra-regional 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 TMRC. 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 TMRC 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.
England there are significant relationships between England and all other parts of Europe except southern Europe and perhaps Fennoscandia (Norway, Sweden, Finland). Norse Vikings therefore left little mark in England, while English Crusaders do not appear responsible for U106 in southern Europe. Curiously, there is a relatively little link between England and the rest of the UK. This northern link with northern Europe (Fennoscandia, Iberia, BaNilea) is clearly present, though significant migrations around the same time fall to link between the English population and central and eastern European populations. Overall, the Anglo-Saxon signature is relatively weak.

Scotland: The Scot-Irish link is both proportionally and numerically the strongest seen in this histograms, showing a very significant migration between the two countries that has been continuous over at least the last 2000 years. This is known to operate in both directions, from the Scot invasions of Scotland to the Plantations in Ireland. Low level links to western Europe over the last 2000 years appear present too. A small peak in migration with central Europe may be present around 2000 years ago, but is difficult to substantiate. There may still be a link between these regions in the far east with central Europe. But the peak is not significant enough to state anything.

Wales: The lack of U106 males in Wales makes it difficult to ascertain whether punitive migrations are real, or statistical artefacts. Links with various parts of Europe around 2000 years ago are marginal, but it is not clear that any of these are substantial.

Ireland: The strong link to Scotland makes it difficult to determine whether links are truly in origin, or whether migration from other parts of Ireland to Ireland was via Scotland. The Scot-Irish link is too strong to resolve this. It is still only a guess at which is going on and what is important. It is one interpretation of a limited set of data. There are many other possibilities.

Migrations are coded into these diagrams as a series of arrows. Grey arrows denote migrations that have happened prior to the SNP. Coloured arrows denote migrations in specific time periods, ranging from blue (last 500 years) to red (2000 years ago or older). Note that, in general, migrators older than about 1700 years quickly become difficult to identify, even when only one marker is taken into account.

A note on histograms:
The series histograms show the distribution of mutations networked in recent common ancestor of a single clade, and the route that their ancestors took to get there. It also shows some of the migrations that their descendents seem to have been involved in. We take this map as a probable scenario. It is only a guess at which is going on and what is important. It is one interpretation of a limited set of data. There are many other possibilities.

**Factsheet: U106**

- **Likely origin**: Germany, 2000 BC
- **Population equivalent association**: Corded Ware Culture, Single Grave Culture
- **Pre-Corded-Four-Border Culture, Bell Beaker Culture**
- **Primary regions**: Germany, Low Countries, north/east France, south/east England
- **Compared to parent, less common in**: Iberia, Italy, Scotland, Wales, Ireland

**Possible migrations**

- **England**
  - Black: 0-360 years → 1557 AD onwards
  - Black: 0-360 years → 1557 AD onwards
- **Wales**
  - Black: 0-360 years → 1557 AD onwards
- **Scotland**
  - Black: 0-360 years → 1557 AD onwards
- **Ireland**
  - Black: 0-360 years → 1557 AD onwards
- **All**
  - Black: 0-360 years → 1557 AD onwards

**About this map**

This map represents the possible place of origin of the common ancestor of men in this clade, and the route that their ancestors took to get there. It also shows some of the migrations that their descendents seem to have been involved in. We take this map as a probable scenario. It is only a guess at which is going on and what is important. It is one interpretation of a limited set of data. There are many other possibilities.