U106 explored: its relationships, geography and history

Report to the U106 group

March 2017 edition

Principal investigator: Iain McDonald

Contents

Methodology & Background

CAVEAT EMPTOR!

This is a very active field of research,. Both testing technologies and the results they are providing are advancing at an impressive rate. New information and methods may be available above those mentioned here. Also, every test and every situation is unique. Different approaches work better for different people, so please treat this document as generalised advice that may not apply directly or most appropriately to your individual situation.

FOREWORD

Our ancestors decisions shaped the world we live in today: whether to fight or flee, what to hunt, who to marry, or simply what to get for breakfast. Countless decisions aggregate to success or failure in life, whether a man survived to produce a family, and ultimately whether they prospered and eventually produced you and I. In this work, we attempt to re-trace some of those decisions back in time to find who our ancestors were, and what decisions were made that allowed them to play a critical role in the history of Europe and the wider world.

Our particular story concerns a large family who carry a particular genetic mutation - worn like a molecular badge - which allows us to identify them as sharing a single common ancestor in which this mutation first arose. That mutation is named U106, or alternatively S21, and is the result of a simple typographical error that happened around 5000 years ago, where one encoding molecule was misread among the 59 million that made up the Y-chromosome of a particular cell. That cell grew into a man, and that man is the 150-times great-grandfather (or thereabouts) of about one in eight men of European descent today.

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

DNA TESTING BASICS

Genetic testing is a comparative science: it compares one person's genes against another person's. There are three main ways of testing DNA: autosomal DNA, mitochondrial DNA and Y-chromosomal DNA.

Autosomal DNA represents the bulk of our DNA, and is inherited in roughly equal quantities from each of our parents. These are the tests commonly used to prove maternity, paternity or very close family relations. Going back several generations, it becomes increasingly more difficult to identify the relative in question: e.g., exactly which one of your 64 great-great-great-great-grandparents you are related through. Also, the exact amount of DNA received from each parent can vary, and the chances of inheriting the minimum size of identifiably similar DNA from any particular ancestor decreases very quickly after a few generations. In small communities, inter-relation can also become a problem after a few generations. This combination of factors means that autosomal DNA starts to lose its efficiency after about 4-6 generations.

Mitchondrial DNA is inherited solely from the mitochondria in a mother's egg cells. Hence, this can be used to trace one's mother's, mother's, mother's, ..., mother. The mutation rate of mitochondrial DNA is relatively slow, due to its small size, hence mitochondrial DNA is largely useful only for solving very specific problems, or looking at a portion of deep ancestry.

Y-chromosomal DNA is inherited solely from one's father. Hence, its most common use is to probe the history of a patriarchal surname: one's father's, father's, father's, father's, ..., father's line. The mutation rate of Y-chromosomal DNA is comparatively quick, but is still slower than one mutation per generation using current and forseeable technologies. The fast mutation rate makes it useful in investigating genealogy over historical times, and building an accurate tree of relationships stretching back millenia. However, the mutation rate (and its uncertainties) still limit exact genealogical work.

This report focusses on Y-DNA testing. Specifically, it focusses on a particular "superfamily", spanning much longer periods than surnames can trace, which represents a large fraction of Europeans and their diaspora today.

DNA is made up of four bases: A, C, G and T, and can be read out as a string of these letters. Genetic genealogy comparing these strings of letters between two or more people's DNA. The differences between them identify mutations that have happened in the transfer of that genetic code from parent to child. These mutations can be used to work out relationships, and the time since their most-recent common ancestor (TMRCA). Associated meta-information, like the locations present in records of people's most-distant known ancestors (MDKA) allows geographical migrations to be uncovered. The combined geospatial and temporal dataset can be used to map out the migrations of our ancestors: when they happened and where they went. Tying these in to known migrations, archaeological cultures, and historical and geophysical events can give some indications of how they lived and why they moved.

Y-DNA TESTING METHODS

There are a variety of different Y-DNA testing options available on the market today. Despite a vast range in prices and performance, they boil down to two different testing methods: SNP and STR testing. As material is passed down, parts of the code can be inserted:

ATGCTGATCGC → ATGCTGATAGATCGC,

deleted:

ATGCTGATAGATCGC → ATGCTGATCGC,

or mutated:

ATGCAGATCGC → ATGCTGATCGC.

The latter kind of mutation, where a single base pair is changed, is called a single nucleotide polymorphism, or SNP (pronounced “snip”).

The other kind of mutation that is frequently used are STRs (Short Tandem Repeats). STR tests are usually performed as a standard set, e.g. Y-12, Y-25, Y-37, Y-67 or Y-111 at Family Tree DNA. These markers take the form of a short section of DNA that repeats a certain number of times. Mutations can cause this number to increase or decrease. A hypothetical example would be:

DYS1234 = 4 TACATACATACATACA

which could mutate to:

DYS1234 = 5 TACATACATACATACATACA

by gaining a repeat.

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

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

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

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

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

General testing advice

WHICH DNA TEST SHOULD I TAKE?

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. Each person has their own ancestors, who followed their own journey. They may be part of a well-populated tree, or a tree that has barely hung onto existence for thousands of years. They may have a lot of distant cousins interested in testing, or they may be all alone.

The combination of factors makes it difficult to provide a one-size-fits-all strategy to answer this question. Generally, 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. However, the details of what you should do depends on exactly what you want to find out.

IDENTIFY YOUR QUESTION

The most important question you need to identify is: what problem you want to solve with DNA testing? Do you want to 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 you want to find out what your deep prehistoric roots are? Different strategies are needed to probe these different questions.

CAN YOU TEST YOUR QUESTION?

Once you have decided on your question, you then need to determine whether DNA can solve it. A common sales tactic for some companies is to advertise that DNA testing can be used to "extend" your family tree, the implication being that it will identify those hard-to-find relatives people refer to as their "brick walls". DNA testing gives information on relationships, but to extend your family tree back you will still need the paper trail research to back things up: DNA testing then becomes a tool in genealogy, not a substitute for paper trail research. If the paper trail simply does not exist, you can forget about breaking through that genealogical brick wall. If the paper trail does exist, you will still have a difficult road ahead.

There are two important steps in moving from the world of paper trails to DNA. Firstly, DNA doesn't lie. It doesn't rely on people keeping accurate paper trails, and it uncovers all manner of illigitimacies, adoptions, extra-marital affairs, and genealogical and mythological mistakes that you never knew existed. Be prepared. That 19th Century book on family history you've been referring to? Probably not accurate. That anciently accepted origin for your surname? Probably only applies to a small fraction of the people with that name.

Secondly, the information it returns is nebulous and imprecise. DNA testing will show that two people are related but (except for very close relations) will not tell you exactly how. You may be lucky if you can identify the millennium of your relationship to someone else in some cases, unless you test very deeply, and you will be very lucky to identify a specific century century. Identifying a specific person is normally only possible with a bucketload of testing of many people, costing well over $1000, and even then it only works if you have a paper trail to back it up.

CHOOSING YOUR FIRST TEST

Assuming your question can be addressed by DNA, you then have to choose your first test.

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 engaging with the people to whom you are most closely related.

To start off with, most people will want to take a moderately deep test. An cheap chip-based SNP test (e.g. National Geographic 2.0 or Britain's DNA Chromo2) will give you an immediate answer as to which part of the family tree you belong to. But it will probably give you zero information about what has happened in the last 1000 years (if you're lucky) or the last 5000 years (if you're not). Conversely, an STR test (e.g. Family Tree DNA's Y-67 test, or equivalent at YSeq.net) will tell you exactly how many people you match and who they are. However, if you don't have any close matches, you may find yourself in a worse situation than if you took an SNP test. It's still the case that most people take an STR test, then follow up with SNP-based testing.

If you choose an STR test, you should maximise the number of markers you can test. There is little point these days in testing 37 markers or less unless you have prior knowledge that you might match someone who has already tested. At least 67 markers are needed to differentiate between the major haplogroups. If money is no object, 111 STR markers are preferable. Do remember though, that your first test is unlikely to be your last - budget wisely!

WHICH TESTING COMPANY?

Each company provides a different user experience:
(1) Family Tree DNA is the most popular: it is widely used for STR tests and next-generation sequencing (NGS: BigY). It has the largest database, which is very important for making comparisons, and project groups which allow users to compare results.
(2) YSeq.net is the cheapest. Its strength lies in cheap, single SNP tests, so is usually for later in the user experience. It also offers cheap STR and whole-genome tests.
(3) Full Genomes Corp. is the most advanced. Its strength lies in the deepest Y-DNA and whole genome NGS tests, so are typically for advanced users.
(4) Britain's DNA offer the Chromo2 test, which is a good, simple, SNP chip test.
(5) Ancestry no longer provide Y-DNA testing.
Of these, only Family Tree DNA invests heavily in user communities and comparisons. For the other companies, it is left to the user to seek out their matches in the online community.

JOINING PROJECT GROUPS

Project groups, either within Family Tree DNA or the wider community, allow individual testers to share their data. External groups allow users to share their data, regardless of the country of origin. Within this group, there will be the expertise to properly analyse that data and make recommendations to individuals about the best steps to take next. I recommend joining at least one haplogroup project, geographical project and surname project.

The R-U106/S21 group that I help administrate is a haplogroup project for R-U106. By collecting data on R-U106 testers together, we provide a sample size greater than almost every professional study (even though it is not so homogeneously sampled as such studies). Perhaps 3% of human male lines are R-U106, so 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. Since DNA is a comparative science, the more data you share with the community, the more information you are likely to receive back about your family history. It is important that you share your data as widely as you feel comfortable with.

SITUATION-SPECIFIC TESTING ADVICE

The following information is aimed specifically at the R-U106 haplogroup, but can usually be generalised to any given haplogroup, particularly those of European origin.

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 and track their migration over the last few millenia. 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: normally the professionals focus on the large-scale structure, whereas normally our interests are typically more specialised. If this happens to be your primary interest, you are going to want to spend fractionally more on next-SNP testing, and invest in next-generation sequencing tests.

Step 1: Basic testing. To start with, you will probably want to find out what has been done on your surname already. A basic test, like a Y-37 test from Family Tree DNA, will be able to tell you if you match anyone else with a variant of your surname. A close match (genetic distance <= 2/37) should normally share your surname's origin. If existing surname matches have already taken a next-generation sequencing (NGS) test, you can use their data. If not, you will have to take your own NGS test.

Step 2: SNP testing. Prehistoric migrations are traced by grouping together people with the same SNP mutations to form new branches of the human haplotree, and comparing their geographies. If you can, save up for an NGS test like BigY, YElite or emerging long-read ("third generation") technologies. These will uncover the SNPs unique to 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 tested, 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, especially those in the majority group R-L11/P311, any of these tests will probably take you what is happening in the period between 5000 years ago and about 1500 years ago. Some people are lucky, and have well-populated 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 "singletons". You want to find people who share some of these singletons with you.)

Step 3: 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. Only a minority of people have taken NGS tests. If you have not done so already, upgrade to 67 or (preferably) 111 Y-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, Semargl.me, or within your haplogroup project.

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 4: encourage your close matches to upgrade. If you have taken a NGS test (BigY/YElite/etc.), you should also ensure your closest NGS 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. If you're very lucky, you and your haplogroup project administrators may be able to work together to identify an origin for your clade. If not, you will need to have your close genetic matches upgrade. Hence, from hereon, I shall assume you have taken an NGS test.

You should then encourage your close STR matches to upgrade to an NGS test. They will hopefully share some of your singletons, and give you a new "terminal" SNP. This will allow a more precise tree structure to be generated, and better ages to be estimated from your data. 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.

During this process, you will need to work with your haplogroup administrators to identify the age of each clade-forming branch in your tree, identifying changes in the geographical distribution of its members. Changes in geographical distribution identify migrations, and tying computed ages to historical or archaeological migrations will tell you the path your ancestors took. Do remember that this can be a controversial process without definitive answers: you are looking to build up a body of evidence, not find unequivocal proof.

HOW DID YOUR ANCESTORS ARRIVE IN BRITAIN?

This can be generalised to any post-Roman migration, although continental European testers will find it more difficult to find suitable matches.

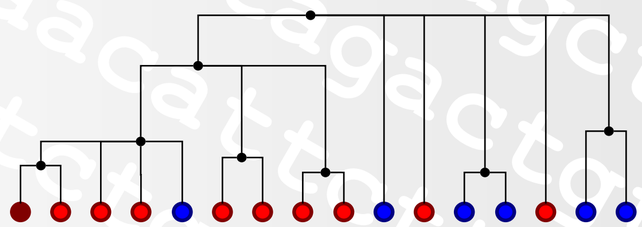
The same advice applies here: start with a basic test, establish what level of SNP testing you need to undertake, then upgrade your STR markers, then encourage your matches to upgrade. Again, you will probably not find a definitive answer, but you can build up a substantial body of evidence.

Unfortunately, not everyone can currently be successful in this quest. Most of the U106 in the British Isles probably arrived here between 800 and 2100 years ago. Since typically NGS testing will find matches only down to timescales of 1000 to 2000 years ago, this means we are working very close to the limit of what can be achieved with our current testing pool. Many people will find that they need some time for more testers to arrive, in order to get a statistically significant sample. It is quite common for small clades of a few people that we think of as "anciently British" to find one tester elsewhere which indicates they could have arrived with the Normans, Saxons, Vikings, or someone else. A typical tester may not expect to get these answers until (at current projections) around 2020.

Step 4a: building a tree. Assuming you come from a well-populated clade, the key factor we need to determine here is the change in geography that signals your family's arrival in the British Isles. This can be complicated by the "founder effect": the presence of small sub-clades (large, recent family groups) that have spread from a recent common ancestor. It's important to try to draw a family tree of all your clade's members to try to work out the geography of each branch. If you can reduce clear branches down to a single representative geography (e.g. a British branch, a French branch, a German branch), then you have more chance of success. This can be a tricky thing to get right: this approach won't necessarily work well using simple estimates like genetic distance. One has to get to know the raw, numerical Y-STR data and be comfortable deciding the order in which mutations formed.

If you can do this, hopefully you will find a point at which the geography of your clade switches. It may be that, somewhere along your line, you will find a clade where (say) six of the sub-clades are British, two are German and one is Swedish; while at the same time, your particular sub-clade has five sub-clades that are British and one from somewhere else in Europe. This could indiciate that the migration to Britain happened at this point, and this is the clade you need to target.

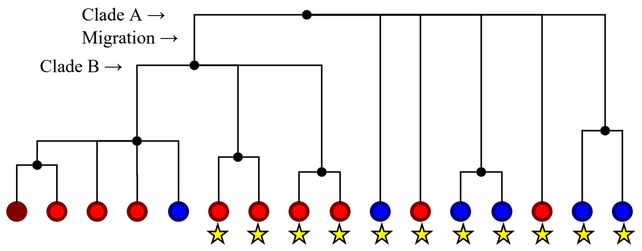
An example of a change from a mostly "blue" geography to a mostly "red" geography. These diagrams can be constructed from Y-STR or NGS results. Each black dot represents an SNP or STR mutation. Each red or blue dot represents a tester from the "red" or "blue" country. The protagonist in our example is the dark red tester.



You need to start by working out your closest STR and SNP matches from continental Europe. You need to find phylogenically 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.

For our previous example, the following chart gives you an indication of the people you need to encourage to upgrade (starred):



Remember that, 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 distantly related tester to cast significant doubt on the "Britishness" of any SNP. It is likely to be several years before we can routinely identify migrations that occurred 1000 years ago.

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

If you want to origins this recent, you obviously have two more pieces of information: your surname, and your Y-STR matches. This means you can think a bit differently.

The same caution holds as before: if you are looking for named ancestors, you will find this difficult. Even with experimental data from the most advanced tests, the resolution possible with these tests still is not as short as a generation. Hence, there is no guarantee that you will be able to isolate a single ancestor from your DNA. Even if you can, translating this into a named ancestor still requires a paper trail to have survived somewhere. For example, finding the father of your most-distant known ancestor is exceptionally difficult, and may not be possible for everyone.

What is slightly easier is to prove descent from someone. This requires triangulation. A good example are the Cheshire families of Warburton and Dutton. Historical records show that the Warburton family descends from a cadet branch of the Dutton family, originating around 1200 AD. The family ultimately descends from the Norman knight, Odard de Dutton. Officially, no branches of either family still exist, however both family names still survive and their members have taken DNA tests. The genetics of both families show several different origins for their surnames. However, a branch of the Dutton family are Y-STR matches to a branch of the Warburton family. Both families have taken Y-111 and BigY tests, and the time to most-recent common ancestor (TMRCA) for their relationship is of the order of 800 years ago. This is highly unlikely to be by chance, and can be considered proof that these families represent the true descent of Odard de Dutton.

For the purposes of argument here, I will presume that your ancestors originate from the British Isles, but the same basic principles can be applied to other locations where western European surnames are used. I will also assume that you have either tested through Family Tree DNA, or are comfortable with accessing their public databases and making the comparisons yourself.

Step 1: Get in contact with a surname project. Some surname projects are better than others: some are run by highly dedicated individuals who know a lot about DNA; some are run by people who have no idea what they are doing but since no-one else has stepped in, they're left with the task; and some were set up by people who have since died and now exist either as ghost projects or run by people with no personal interest in the surname. Hopefully you are lucky, because usually your surname project administrator will know more about your surname and the research that's been done already than you do. If you already know roughly where your ancestors come from, you will also want to join a geographical project.

Most surnames have a diverse range of origins. Almost without exception, not everyone with the same surname will be related to each other. It is quite common to find surnames with a dozen different origins, or even hundreds. This is a particular problem in the Scottish clan system, and in Scandinavian countries that used patronyms until recently. But it also occurs in places with inherited surname patterns more typical of western Europe. There are a variety of reasons why this is: most people's first conclusion is cuckoldry, but this only happens at about 1-2% per generation. Many surname changes have occurred by people wishing to distance themselves from their families, or from the law, because of adoptions of orphans, or simply to stake their claim on a new place. Either way, it is rare for a surname project to be able to tell you the person you originate from without some genetic testing.

Step 2: Basic testing. Because you are looking at recent relations, you will need a higher resolution test. You will probably want some combination of STR and later SNP tests: the SNP tests will tell you which branch of the family tree you need to look at, while the STR tests will give you the information you need to match people closer to the present day. If you are looking at medieval, surname-level relationships, you don't need the ultimate high-resolution, but you will need some reasonably high level of testing. You may want to start with a 67-marker STR test, since a large fraction of people have now bought this level of testing.

Step 3: Find your STR matches. For Family Tree DNA customers, these are automatically given to you. For other customers, you may have to use tools like YSearch.org and/or semargl.me to identify close matches. If you are lucky, you will have a lot of matches who share your surname, and they will have already done a lot of your research for you. If you have a lot of matches and they are still floundering in the dust, you may have to take charge! Of course, you may have no surname matches at all, which will make your life much more difficult. And you may have a lot of matches from a different surname, in which case you might want to start wondering if your surname has changed at some point.

Step 4: Take an SNP test. This goes back to the concept of placing brackets in time around the individual you want to study: in this case the origin of your surname. The best SNP test depends on your budget and your matches: I'm assuming that, if you're in a well-managed surname project with lots of matches, they can give you more specific advice, or at least help you best apply this advice to your situation. Ideally, you will want to take a next-generation sequencing test (e.g. BigY, YElite, or WGS test). An NGS test will give you the SNPs that have occurred in your recent past: shared SNPs, and "singleton" SNP that are private to your test. It is becoming increasingly important for at least one person in a family group to do it, and preferably two or more. The choice of NGS test may be dictated by the test's temporal resolution (see later).

Step 5: Identify geographic origins. Hopefully now you will know everything you need to about your own DNA, and you may have some idea about who and who doesn't match you. If you are very lucky, you will now have an origin in Europe to test around. If you do not, then you will have to find one. The next steps 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 have any additional information (like whether your surname is a toponym), the number of existing matches you have, and how many people you can convince to pay for themselves. The rest of this is therefore even more generalised advice.

The first thing you will need to do is identify where your surname is common, in order to look for "hotspots" of people to pick off. The following website details the distribution by county of most UK surnames:

http://gbnames.publicprofiler.org/Surnames.aspx

Similar websites exist for Germany and some other countries. Using this technique, you can 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 identified a set of likely counties, you may find it beneficial to narrow things down to the parish level. This can often identify a few parishes where your surname has been historically most active, and can break county-level matches up into two or more distinct groups. 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. This parish-level work can be important for common surnames: for example, my own surname (Donald) has at least four origins attested by DNA within a few miles of Aberdeen, which can be localised to individual parishes using the 1696 tax registers.

Step 6: probing "hot spots". The next step is to get DNA from people in these hot spots. If you are lucky, some people will already have taken DNA tests which will let you see if people from these regions 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 other 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. It's important at this point that you do not simply become a salesperson for your chosen DNA company, so you will need to think about the potential pros and cons of each test, in order to give your potential matches a realistic idea of what they will get for their (or your) money. Hopefully, you will have done some kind of NGS test or at least an SNP pack test, which has put you into a relatively recent, relatively rare branch. You can use an SNP from this branch (not from among your singleton SNPs) to test whether people belong to your wider family. For YSeq.net, you should be able to do this for US$17.50 plus postage. This provides a very cheap test that you can farm out to many different potential matches, at their expense or yours. However, this is simply a binary indicator of whether they are related to your or not. 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 act in their interests as well as your own! An example of such a test would be a 37-marker STR test, which will give them a head start on their own genetic pathway.

Step 7: detailed exploration. Once you have covered all the "hot spots" with cheap testing, you should hopefully have some new SNP or STR matches who are related to you, so roughly which of the hotspots your ancestors are likely to have come from. You should then focus on these region(s), and get a few people with your surname to test (and surnames of your closest STR matches if applicable). It's important to contact the people with the longest provable lineages you can find. You should find that some or all of the people in that region who share your surname are from your family.

Step 8: upgrade your matches. Upgrading your matches to Y-67, Y-111 and/or NGS tests will then give you the data you need to bring the branches of your family together. You should then construct a phylogenic family tree, showing how your lines fit together and when each mutation is likely to have occurred. Note that this may require a lot of testing to work. 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.

AUTOSOMAL TESTING AND MORE RECENT MATCHES

For Y-DNA testing, the shorter the timescale you are looking on, the more advanced the test you need. Currently there is no test that can separate individual generations, except in a small percentage of lucky cases, so old-fashioned detective work will be required to supplement your genetic testing.

However, if you are looking to find recent connections, you may find benefit in an autosomal test. Generally, basic autosomal tests currently have the ability to find matches up to about fourth cousins. In certain circumstances, and with careful analysis and a lot of luck, one can sometimes find matches more distantly than this, but the random inheritance of autosomal DNA and potential for inbreeding as you expand your family tree means that tests at this level are less reliable. The ISOGG Wiki gives information on how likely you are to find an autosomal match for a variety of situations:
isogg.org/wiki/Cousin_statistics

With relationships that are slightly older, say in the fourth to tenth cousin range, some extra comparisons may be possible by testing specific autosomal SNPs with whole genome sequencing (WGS) tests. The technology to perform these tests has only recently come to fruition, and I am not aware of any indivdiual studies that have successfully used these tests for this particular purpose, but the principle is there.

IN CONCLUSION

Everyone's situation is different, everyone's desires are different and information available to each individual person is different. This advice gives only the general principle behind a scientific approach to understanding your ancestry. It should be taken with the advice of your appropriate surname and haplogroup project advisors. They may well be able to give a much more direct approach to your person situation. At each step, use your judgement to ask "what is this test going to get me, and is there a better way to go about it".

The rest of this document is dedicated to helping users decide between different tests, and the depth of testing that they can provide.

CHOOSING BETWEEN TESTS

The rest of this document is dedicated to helping users decide between different tests, and the depth of testing that they can provide. An up-to-date comparison can be found on the ISOGG Wiki page:

http://isogg.org/wiki/Y_chromosome_DNA_tests

The ISOGG Wiki gives the relevant statistics and pricing points for each test. Note that many companies (most notably Family Tree DNA and, less commonly, Full Genomes Corp.) will offer substantial sales on these prices). These can often be worth waiting for, rather than taking the plunge immediately.

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: YElite is about 35% 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 haplotree 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.

Full Genomes Corp. 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 results. It's a small thing, 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 let you know which singletons you truly share with other people (the U106 group does this for free!). YFull will also give you an age (although for U106 they have fewer testers so it will be less accurate).

Step 6: probing "hot spots". The next step is to get DNA from people in these hot spots. If you are lucky, some people will already have taken DNA tests which will let you see if people from these regions 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 other 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. It's important at this point that you do not simply become a salesperson for your chosen DNA company, so you will need to think about the potential pros and cons of each test, in order to give your potential matches a realistic idea of what they will get for their (or your) money. Hopefully, you will have done some kind of NGS test or at least an SNP pack test, which has put you into a relatively recent, relatively rare branch. You can use an SNP from this branch (not from among your singleton SNPs) to test whether people belong to your wider family. For YSeq.net, you should be able to do this for US$17.50 plus postage. This provides a very cheap test that you can farm out to many different potential matches, at their expense or yours. However, this is simply a binary indicator of whether they are related to your or not. 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 act in their interests as well as your own! An example of such a test would be a 37-marker STR test, which will give them a head start on their own genetic pathway.

Step 7: detailed exploration. Once you have covered all the "hot spots" with cheap testing, you should hopefully have some new SNP or STR matches who are related to you, so roughly which of the hotspots your ancestors are likely to have come from. You should then focus on these region(s), and get a few people with your surname to test (and surnames of your closest STR matches if applicable). It's important to contact the people with the longest provable lineages you can find. You should find that some or all of the people in that region who share your surname are from your family.

Step 8: upgrade your matches. Upgrading your matches to Y-67, Y-111 and/or NGS tests will then give you the data you need to bring the branches of your family together. You should then construct a phylogenic family tree, showing how your lines fit together and when each mutation is likely to have occurred. Note that this may require a lot of testing to work. 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.

AUTOSOMAL TESTING AND MORE RECENT MATCHES

For Y-DNA testing, the shorter the timescale you are looking on, the more advanced the test you need. Currently there is no test that can separate individual generations, except in a small percentage of lucky cases, so old-fashioned detective work will be required to supplement your genetic testing.

However, if you are looking to find recent connections, you may find benefit in an autosomal test. Generally, basic autosomal tests currently have the ability to find matches up to about fourth cousins. In certain circumstances, and with careful analysis and a lot of luck, one can sometimes find matches more distantly than this, but the random inheritance of autosomal DNA and potential for inbreeding as you expand your family tree means that tests at this level are less reliable. The ISOGG Wiki gives information on how likely you are to find an autosomal match for a variety of situations:
isogg.org/wiki/Cousin_statistics

With relationships that are slightly older, say in the fourth to tenth cousin range, some extra comparisons may be possible by testing specific autosomal SNPs with whole genome sequencing (WGS) tests. The technology to perform these tests has only recently come to fruition, and I am not aware of any indivdiual studies that have successfully used these tests for this particular purpose, but the principle is there.

IN CONCLUSION

Everyone's situation is different, everyone's desires are different and information available to each individual person is different. This advice gives only the general principle behind a scientific approach to understanding your ancestry. It should be taken with the advice of your appropriate surname and haplogroup project advisors. They may well be able to give a much more direct approach to your person situation. At each step, use your judgement to ask "what is this test going to get me, and is there a better way to go about it".

The rest of this document is dedicated to helping users decide between different tests, and the depth of testing that they can provide.

CHOOSING BETWEEN TESTS

The rest of this document is dedicated to helping users decide between different tests, and the depth of testing that they can provide. An up-to-date comparison can be found on the ISOGG Wiki page:

http://isogg.org/wiki/Y_chromosome_DNA_tests

In choosing a test, the two main factors you need to look at are the quality and quantity of the final information that it gives you about your ancestry. Note that this can be quite different from the quality and quantity of each test.

Inherent quality: bespoke STR tests and bespoke SNP tests (including sequencing tests) provide highly accurate results. STR data can be retrieved from sequencing tests, but at a lower reliability. You should not anticipate extracting accurate STR data from a sequencing test like BigY or YElite: data for the shorter STRs will be reliable, but longer STRs will increasingly not be. As technology moves to longer reads (e.g. tests trialed using Oxford Nanopore) then more STR data should be extractable.

Reproducability: reproducability can be an important factor. If you test a 111-marker STR panel twice on the same person, you should expect exactly the same results. If you test a SNP twice, you expect the same result. By contrast, some chip-based tests and most sequencing tests don't have 100% reliability, and some information is provided on a best-effort basis. For example, a BigY test typically covers bewteen 10 and 11 million base pairs of DNA, yet two BigY tests will typically only have 98% overlap. The remaining 2% of positions will not be accurately called in the other test. The can provide some ambiguity about whether certain SNPs are shared or not.

Derived data quality: the difference in mutation rate between STRs and SNPs becomes important when we try to derive relationships. STRs will typically mutate once every few thousand or tens of thousands of years. SNPs will typically mutate a little slower than once every billion years, although there are variations in both. This causes problems when making trees from STR results, because they can mutate both forwards and backwards: you don't always know if, say, an STR has mutated from an allele of 13 to 14 and then back to 13 again, or if two people with an allele of 14 are as a result of the same mutation or different mutations. For that reason, SNPs work better for making trees than STRs.

Inherent data volume: a larger test means a better, more certain, more precise match, so a Y-111 STR test will be better than a Y-67 test, which will be better than a Y-37 test, because there will be more STR mutations that can be matched between people to find genuine matches and remove chance matches. Equally, a 14-million-base YElite test will be better than a 10-million-base BigY test, because it will discover more SNPs. However, the test volume is only useful if you are comparing people who have taken a similar level of testing: e.g. comparing your Y-111 test against someone else's Y-67 test won't get you any better results than comparing your Y-67 test against someone else's Y-67 test. The same is true of comparing YElite versus BigY tests instead of BigY versus BigY tests. Conversely, someone has to be first to take a new test, and we usually find that one person testing in a group is good encouragement to others. On balance, the ideal situation is to co-ordinate your tests and upgrades with other members of your group.

Number of matching tests for comparison: testing for large numbers of SNPs, e.g. via sequencing tests, is a relatively recent and expensive technology. Consequently, most people have still only taken STR tests: perhaps only around 1% of genetic testers have taken any deep form of seqeuncing test. Since genetic genealogy is a comparative science, it's important to have a large number of matches. For that reasons, STRs work better for finding matches than SNPs. Consequently, most new testers still opt for a combination of STRs and SNPs, with the exact tests chosen depending on their budget, genealogical situation and problem.

Analysis: many haplogroup projects are now set up to handle sequencing data from Family Tree DNA's BigY tests and Full Genomes Corp.'s YElite and Whole Genome Sequencing tests. However, many groups rely partly or wholly on BAM file interpretation services provided by YFull.com, Full Genomes and others, which cost around $50. Before committing to an expensive sequencing test, it is important to determine what your haplogroup project is capable of doing, and account for the extra expenditure as necessary. The U106 group is currently (March 2017) set up to process both BigY and Full Genomes' test results, but the matching between them is currently done manually. As a general rule, it is advisable to determine which tests your closest matches have taken, and try to order at least the same level of testing if possible. This ensures the maximum amount of both of your tests will be useful. Note that each company maintains its own separate database of results, so results from one company will not become part of another's database: e.g. Full Genomes or Yseq results will not be transferrable to Family Tree DNA.

Data value for money: Particularly for next-generation tests, value for money is important, but determining value for money can be a tricky subject. You need to get more SNPs to split up a group to see where the branches form in the family tree (this can be especially useful for recent families), and need more coverage to refine the time of a relationship. Doubling the coverage of a test doubles the number of SNPs you will find. However, measuring the relationship time depends on the square root of the coverage, so quadrupling the coverage of a test halves the uncertainty on a relationship. So going from the BigY test (10 Mbp) to YElite (14 Mbp) to upcoming "third-generation" tests (about 20 Mbp) increases the number of SNPs by 40% for YElite and 100% for the 3G tests, but it will only improve the time estimate by reducing the uncertainties by 15% and 30% of what they were. Whether this justifies the extra 30% price for YElite (or more for 3G tests) depends on what you want to do with them and who you have to compare to. Most people want both to find the branches and get a better time estimate, so on balance YElite is marginally better value than BigY... except when BigY is on sale!

An alternative to paying more for higher test coverage is to buy two lower-resolution tests of distantly related people, for the price of a single higher-resolution test. An application of this might be buying a BigY test for yourself, and another BigY test for one of your closest matches, instead of taking a third-generation test. Depending on your aims (see above), this may be a more attractive option if you are interested in either getting better time resolution. As always, every circumstance is different, and for certain applications like winkling out associations between ancient clades and getting the finest time resolution on the last 15 or so generations, the highest-resolution tests may be necessary.

TIME RESOLUTION

The following comparisons give the temporal accuracy you can expect from each test.

Y-12

GD = 0 : 0 - 2130 years
GD = 1 : 150 - 3360 years
GD = 2 : 390 - 4590 years

All things being equal, for any given match you have with a genetic distance of zero at 12 markers, 95% of the time you can expect them to be related at some point in the last 2130 years. If you have a match at a genetic distance of one of out 12, you can expect them to be related to you in the last 150 to 3360 years. At a genetic distance of 2/12, you will normally be looking at a relationship in the last 390 to 4590 years. As you can see, a 12-marker test is not often normally very informative except to define broad haplogroups. Today, they are normally only used to confirm relationships.

Another way of looking at this is that a Y-12 test will normally identify almost everyone who is related to you in the last 150 years, most of the people related to you in the last 1500 years, and a fraction of the people related to you in the last 3360 years.

Note that, whereas you might only match a handful of men in your family in the last 150 years, most people will match many millions of men within the last 3360 years. So the distribution of your matches will be skewed to the older end of these ranges.

The end result of this skewed distribution is that contamination from more distant matches is important. There are probably around 100,000 P312 tests in Family Tree DNA's database. To estimate contamination, we need the average number of mutations.

To calculate this, I have taken a large dataset of Y-STR tests (around 500) from the clade DF98, which is about 4500 years old. It is sufficiently close to U106 (about 5000 years old) that the results should be fairly representative for the whole of U106. In this clade, there are an average of 2.17 mutations in the last 4500 years: some people have none, some people have up to four. Each of the 12 STRs can mutate up or down, giving 24 possible mutations. Making the simplifying assumption that two mutations is typical, there are 24 x 22 = 528 possible combinations. For a U106 person, the chances are that a couple of hundred P312 people (100,000 / 528) will have had the same 12-marker mutations as you have. This means that the Y-12 matches of almost everyone are pretty useless at determining whether a relationship is within the last 5000 years or so. In many cases, a Y-12 match and a matching surname will indicate a relationship, but for very common surnames even this isn't a given. For this reason, Y-12 tests have generally become "specialist" orders.

Y-25

GD = 0 : 0 - 750 years
GD = 1 : 60 - 1170 years
GD = 2 : 120 - 1530 years
GD = 3 : 240 - 1890 years
GD = 4 : 360 - 2280 years
GD = 5 : 510 - 2640 years

The DF98 sample shows an average of 5.28 mutations in each test. However, as many as one in 200 people can still be expected to have no mutations in the first 25 markers over this 4500-year timeframe, and one in 30 can expect to have only one. A Y-25 test gives 50 possible mutations, though for multi-copy markers like DYS464 these aren't always individually differentiable. Five mutations therefore gives about 50 * 48 * 46 * 44 * 42 ~ 200 million possible combinations, although this is normally restricted to many fewer, especially as some markers mutate much faster than others. Most people, but not everyone, can therefore be secure that their exact 25-marker matches are related to them in the last couple of millennia. A small fraction of families will be lucky enough to have a set of mutations unique to their family. However, as many as 3% of people will not be able to differentiate major clades which haave happened in the last 5000 years, even for exact 25-marker matches. Hence, Y-25 tests have generally either been discontinued or have become "specialist" orders.

Y-37

GD = 0 : 0 - 330 years
GD = 1 : 30 - 570 years
GD = 2 : 60 - 660 years
GD = 3 : 90 - 840 years
GD = 4 : 150 - 990 years
GD = 5 : 210 - 1140 years
GD = 6 : 270 - 1290 years

In the DF98 sample, we observe an average of 9.49 mutations. There should be extremely few people who have no Y-37 mutations in the last 4500 years, but a very small percentage of people might only have one, two, or three. As many as one in six people will experience some spillover from 5000-year-old clades at a genetic distance of 4/37, but the majority of people should find that their Y-37 matches are from within the last 1000-3000 years, depending on their ancestral population. In the absence of prior knowledge, this makes the Y-37 test the cheapest that will provide any meaningful information on the ancestry of first-time testers. It has the resolution to determine whether anyone else of your surname matches you, and give you some first ideas about where your more distant ancestry comes from. If you aren't wanting to spend a lot of money on a Y-chromosome sequencing test, it almost universally gives you enough information to determine your haplogroup, allowing you to undertake further SNP testing cheaply. I typically recommend either Y-37 or Y-67 for new users.

Y-67

GD = 0 : 0 - 270 years
GD = 1 : 0 - 480 years
GD = 2 : 30 - 51 years
GD = 3 : 60 - 630 years
GD = 4 : 120 - 750 years
GD = 5 : 150 - 840 years
GD = 6 : 210 - 960 years
GD = 7 : 240 - 1080 years
GD = 8 : 300 - 1170 years
GD = 9 : 360 - 1290 years

A Y-67 test will typically identify your relations within the last 1000 years, but there will be contamination from less-closely individuals. In some very populous clades, distant matches at GD = 6/67 may extend back to 2000 years ago or a little more. In about 2% of family lines, matches at 6/67 may end up stretching as far back as 5000 years, but closer matches will certainly be family. However, with 13 mutations on average, there is normally enough information to clearly identify family branches, and structure within them. It allows many families to explore their structure and relationships in ways the Y-37 test can't. This is a good test for first-time testers, and the one I recommend if people can afford it. As the above age ranges show, it is the first test that can realistically separate recent relationships with some degree of rigour. The 67-marker panel also includes a number of slower mutations, which are very useful for placing people in clades with known STR motifs (sets of common mutations). Specifically for U106, it also contains the DYS492 STR, which is a very good marker of being U106 (97% effective). It is almost guaranteed to find everyone who has tested who is related to you within the last 360 years, about half of tested people related to you in the last 700 years, and a smaller percentage of people related to you on timescale of more than 1290 years, making it a good tool for identify recent matches. However, remember that you will not necessarily be able to differentiate someone related a few centuries ago from someone related to you 1000 years ago with good confidence.

Y-111

GD = 0 : 0 - 150 years
GD = 1 : 0 - 150 years
GD = 2 : 30 - 330 years
GD = 3 : 30 - 390 years
GD = 4 : 60 - 450 years
GD = 5 : 90 - 540 years
GD = 6 : 120 - 600 years
GD = 7 : 150 - 660 years
GD = 8 : 180 - 720 years
GD = 9 : 210 - 780 years
GD = 10 : 240 - 840 years
GD = 11 : 270 - 900 years

A Y-111 test is the largest commercially available STR test, although sequencing tests will uncover many STRs and STR-like features. Typically, a Y-111 test will provide a fairly clean set of people who are related to you within the last 1000 years or so. With a typical 22 mutations in the medium-sized haplogroup of DF98, a Y-111 test normally provides a good motif for pinning down your location in the family tree very well: over 99% of people should have at least 11 mutations within the last 5000 years. For people with matches who have SNP tested already, it is normally making it possible to associate them with a recent clade or even an individual family on the basis of a Y-111 test. However, only a fairly small fraction of people have taken a Y-111 test, making matches more difficult to find. If both testers have Y-111, it is a very useful way of separating distant from close relationships.

The Y-111 tests are very useful for working out how closely people are related: they are effective over the few-centuries timescale of surnames. While they can't nominally separate people who are related 270 years ago from people who are related 900 years ago, multiple testers from within a family can start to put better constraints on relationships, normally getting them to well within a factor of two uncertainty overall.

SEQUENCING TESTS

Sequencing tests may be purely Y-chromosome, or they may be ``whole genome sequencing" (WGS) tests that recover the mitochondrial DNA, autosomal chromosomes and even the exome too. Several companies are now offering these tests. The leading contenders in the market are Family Tree DNA's BigY tests; several tests by Full Genomes Corp., including YElite and WGS tests; and the WGS test by YSeq.

For these tests, the key criterion is the callable loci: the total number of base pairs of DNA where a positive or negative call can be made for an SNP. The quality needed to make a call varies between companies. For example (see overleaf), Family Tree DNA call 10.6 million base pairs in their BigY test, but Full Genomes Corp. only cover 8.8 million base pairs when they analyse the same BigY tests. A comparative chart is made below. Note that test data for this comparison were not available at the time of writing for the YSeq WGS test, or the "third-generation" (3G) FGC long-read pilot test.

An up-to-date comparison chart can be found here:

https://isogg.org/wiki/Y-DNA_next_generation_sequencing

but the table below summarises some various properties. These are based on a mutation rate of around 0.82 SNP mutations per billion base pairs per year.

Company	**FTDNA**	**FGC**	**YSeq**	**FGC**	**FGC**
Test type	**Y**	**Y**	**WGS**	**WGS**	**WGS**
Test name	**BigY**	**YElite2**	**15x**	**30x**	**"3G"**
Coverage	8.8	14.8	~13.2	~14.9	~20? Mbp
Years/SNP	139	82	93	82	61

The number of years per SNP gives the shortest meaningful time frame that can be explored using that sequencing test. SNPs occur randomly, so not every test will have an SNP forming within this time frame. The probability of finding one or more SNP sin each test scales roughly as follows:

Years	**BigY**	**YElite2**	**15x**	**30x**	**"3G"**
30	19.4	30.5	27.6	30.8	38.8 %
60	35.1	51.7	47.5	52.1	62.6 %
90	47.7	66.4	62.0	66.9	77.1 %
120	57.8	76.6	72.5	77.1	86.0 %
150	66.0	83.8	80.1	84.1	91.4 %
180	72.6	88.7	85.6	89.0	94.8 %
210	77.9	92.2	89.5	92.4	96.8 %
240	82.2	94.5	92.4	94.7	98.0 %
270	85.7	96.2	94.5	96.4	98.8 %
300	88.4	97.4	96.0	97.5	99.3 %
330	90.7	98.2	97.1	98.3	99.6 %

Any study wanting to reconstruct a family tree based solely or mostly on DNA data requires at least one mutation to occur in every generation, so needs an average mutation rate substantially faster than this. No test can currently reach one mutation in every generation, although current tests can offer one mutation (STR, SNP, or indel) every ~3 generations, and forseeable tests can offer one mutation every ~2 generations.

This means that current and forseen technology can provide relatively accurate brush-strokes to fill in a family tree, but will not normally be able to provide the detail required to create an exact family tree without the autosomal DNA and/or paper trails to back it up.

EXTRACtING STRs FROM SEQUENCING TESTS

Extraction of STRs from sequencing tests is possible, thereby nominally saving the tester from the expense of a bespoke STR tests. However, users should be warned that only a fraction of STRs can be recovered from sequencing tests accurately. This varies between around 80 and 107 STRs, depending on the test, and typically the reliability of these calls is less than it is for bespoke STR tests. The critical factor is the length of each sequence of data that is read: the read length of the test. Current tests cannot reach the read length of several hundred needed to sequence every STR, although "third-generation" tests are being trialled with read lengths considerably over 150 base pairs which should provide this level of reliability in the near future.

Overview of the Y chromosome

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

Source material:
H. Skaletsky, et al., 2003, *Nature*, **423**, 825
P. Francalacci, et al., 2013, *Science*, **341**, 565
B. Trombetta, et al., 2014, *Mol. Biol. Evol.*, **31**, 2108
Selected loci from: ybrowse.org

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 59,373,566 base pairs. (The newer Build 38 is slightly shorter, as several gaps have had their sizes changed.)

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/YElite), 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.

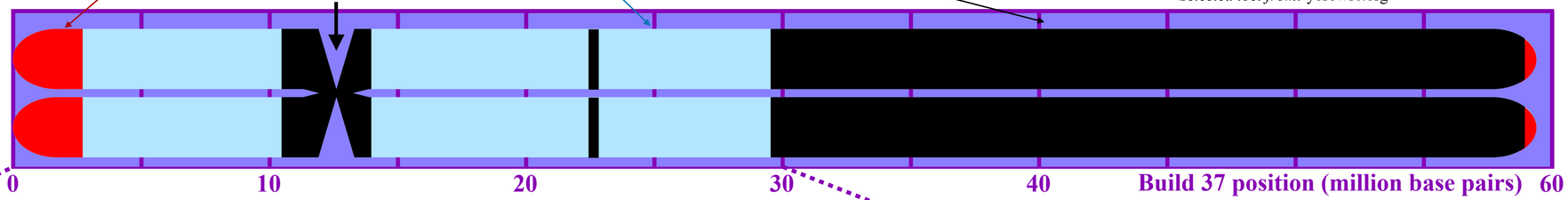
This process requires knowing: (a) which chromosome you are looking at and (b) which place on the chromosome you are looking at, so regions that look like other chromosomes and very repetitive regions usually cannot be read accurately. Only the *euchromatic* regions can.

PSEUDO-AUTOSOMAL REGIONS
These regions are very difficult to read, as their coding strongly overlaps with those of the other (autosomal) chromosomes. These regions are not inherited strictly via the male line, so are not of use to general Y chromosome studies.

CENTOMERE
The crossing point of the chromosome, where the two arms join together.

EUCHROMATIC REGIONS
These are the easily read regions of the Y chromosome where most SNPs and STRs are found.

HETEROCHROMATIC REGIONS
These are regions of very repetitive encoding. This makes these regions very difficult to sequence. 65000 bp of useful information can be found at the end of the long Yp12 heterochromatic region, before the pseudo-autosomal region PAR2. This is not shown on the plots below.



COVERAGE: The claimed coverage of 510 Family Tree DNA's BigY and 8 Full Genome Corp.'s YElite 1.0 tests are shown on the plots below. The companies differ in the quality needed to claim an SNP is called accurately. An intermediate trace is included, showing coverage of 7 BigY raw data (BAM files) as analysed at Full Genome Corp. by Vince Tilroe, using the same procedure as YElite. This allows the two tests to be compared to each other fairly.

There are a number of regions in BigY where a large fraction of the SNP calls are later found to be problematic, an called inconsistently with the structure we find in the rest of the tree. These are labelled. Ignoring these regions increases the fraction of repeatable SNPs from 76% to 87%. Similar regions are not shown for FGC as there are insufficient tests to make a consistency call in a manner that can be fairly compared against BigY.

TYPES OF EUCHROMATIC DNA:

OTHER: Other regions not fitting the above categories.

X-DEGENERATE
Regions with origins predating the split between X and Y chromosomes (~166 million years ago). These regions are largely similar between X and Y, but the accumulated mutations mean they are generally separable.

X-TRANSPOSED
These regions are copies of regions of the X chromosome that are reproduced in the Y chromosome with ~99% repeatability. They are very difficult to sequence due to this similarity.

AMPLICYCLONIC
Regions of DNA that are largely unique to the Y chromosome. These are the easiest to sequence, but contain palindromic regions which are more difficult.

PALINDROMES
Regions of repeating DNA of the form (e.g.) AGCT…TCGA, many of which form palindromic loops, or extensions of the DNA out of the main strands. The multicopy markers lie on these arms, with one marker on each side of the palindrome.