DNA and its Uses in Genealogy

The following is my attempt to provide a simple explanation of DNA and how it can be used in genealogy. It is meant as an introduction. The books and sites in the reference section that follows provide a more professional and detailed explanation.

Part 1: What is DNA?

Think of DNA as a string of letters. There are 4 possible letters - A, C, G and T (these are the initials of the chemical molecules that are represented by the letters).

This string of letters is divided up into chromosomes (so called because scientists use coloured dyes to identify them). Chromosomes in turn contain genes. A gene could be defined as the shortest string of letters that actually does something useful in our development. However between the genes are strings of useless or junk DNA that do nothing (maybe they did once in earlier stages of evolution). These useless strings of letters are important for genealogy but I will come back to that later.

The human genome (i.e. our DNA) consists of 46 chromosomes, or rather 23 pairs. We inherit one set of 23 from our father and one from our mother. When we come to pass on 23 chromosomes to our children we pass a mixture, some from our father and some from our mother. The mixture is different every time.

One of the pairs of chromosomes is the X and Y-chromosomes. A female has 2 X-chromosomes, one from each parent. However a male has an X-chromosome from his mother and a Y-chromosome from his father. It is a gene on the Y-chromosome that causes a baby to be a boy. In the absence of this gene the default is always to produce a girl.

The significance of this is that (unlike all the other chromosomes) the Y-chromosome is never mixed with a copy from the mother. It passes unchanged from father to son through the generations.

Now the body is very good at faithfully copying DNA from generation to generation, but it is not perfect (otherwise evolution wouldn't work). Very occasionally a copying mistake occurs. For example an A may become a T. If it happens in a gene it may cause disease, or rarely it improves the gene. But if it happens in junk DNA it has no effect and so the mistake continues to be copied from generation to generation. It is these differences that make DNA useful in historical and genealogical studies.

There is one other piece of DNA that is passed unchanged from generation to generation. It is in addition to the 46 chromosomes and acts as the energy source for a cell. It is called mitochondria and is only passed down the female line. Males do have it, inherited from their mother, but don't pass it on.

Part 2: Genetic Clans

This section is not specific to genealogy. That will come later. Firstly I want to introduce SNIPs.

I mentioned above that very occasionally a DNA letter is copied wrongly e.g. an A becomes a T. This is known as a Single Nucleotide Polymorphism or SNP or SNIP. So a SNIP is when a letter is copied wrongly.

SNIPs occur very rarely. A specific letter may have only changed once in the whole of modern mankind's existence (150-180,000 years). By concentrating on certain specific sequences of junk DNA researchers have classified the human population into a small number of classes. Of course they have a special name for them i.e. haplogroups, though it is easier to call them clans.

They have done this with both mitochondria, and the Y-chromosome. For mitochondria there are 36 clans, 13 in Africa. There are even fewer Y-chromosome clans.

Ninety-five percent of Europeans fall into 7 mitochondrial clans. Professor Brian Sykes has written a book called **The Seven Daughters of Eve** which gives these 7 clans names. Your clan can be determined by looking at a specific 400 letter sequence of mitochondrial DNA. Just 2 copying mistakes, at positions 69 and 126, define me as being in clan J or Jasmine. I have only one other mistake, at 366, which means I am in the main group of Js, not in one of a number of sub-clans that have been defined.

Researchers do look at other bits of the mitochondria for more definition of the clans, but the 400 letters are all that is needed to position you.

How is this information used? Well firstly it is used to build family trees (again they have a fancy name - phylogenetic trees). These trees show the relationships between the various clans, and within them. It is done by working out the sequence in which SNIPS must occur. For example my SNIP at position 126 is shared with another clan (Tara) but that at 69 is unique to J so must have occurred later.

Perhaps the most startling assertion, though it is logical when you think about it, is that each mitochondrial clan must be descended from a single woman. The copying error occurred just once, so everyone carrying the error must be descended from the first woman to carry the error. Not only that, but by linking the clans in a tree, one clan becomes the source of all the others. Not unnaturally the origin of this clan is called Eve. This doesn't mean that Eve was not one of a population of similar early humans. It's just that no descendants of her contemporaries exist today.

Interestingly, if the same thing is done with the Y-chromosome, 'Y-chromosome Adam' seems to have lived a lot later than Eve (60-80,000 years ago, rather than 150,000 years ago).

Next, by looking at the distribution of the clans in the world, and the amount of change that has happened in various locations, researchers can determine the migrations of man around the world. Of course this has to be linked to the archaeological evidence to be meaningful. Also by doing exactly the same with the Y-chromosome they can build a separate, corroborating picture.

Studies of this can get extremely complicated, and employ lots of probability calculations. The results though can be fascinating. For example my Mitochondria, and my Y-chromosome have completely different histories. Mitochondrial Clan J originated in the Near East and only entered Europe after the last Ice Age with the first wave of farmers. It followed 2 tracks, one of which followed the coast around the Mediterranean, and eventually up to Britain. My Y-chromosome came with the first migrants to come to Europe 40,000 years ago (the Aurignacian culture). They spent the last Ice Age in a refuge near the Pyrenees, where the last mutations took place, and then moved back into Europe as the ice retreated. It is by far the most common clan in Western Europe, particularly along the western seaboard. My results are described in more detail in **Commentary on Results** under **My Mitochondria**.

So to summarise, SNIPs, or letter copying errors are sufficiently infrequent to allow the world population to by classified according to which SNIPS a person has. This is done both for mitochondria, which is passed down the female line, though present in men, and the Y-chromosome, which is passed down the male line and is not present in women. By working out sequences of change the clans can be linked into a family tree, and by looking at the distribution of clans, the amount of subsequent changes in the various localities, and at corroborating archaeological evidence the history of man's migration out of Africa and around the world can be determined.

Part 3: A DNA Test for Genealogists

So far we have discussed SNIPs. Unfortunately these occur too infrequently to help with genealogy. Every related male Warburton would probably have an identical profile. Fortunately there is another test which is more helpful, but it can only be carried out on the Y-chromosome.

It so happens that there are some short DNA sequences that are repeated several times. Whereas with SNIPs we were dealing with a change to a single letter in the DNA sequence, you can think of these sequences as words that are repeated several times. Every now and then the number of copies of the word changes. For example one may be added so, whereas there were 10 repeats before, there are now 11.

These strings of words are called Short Tandem Repeats (STRs), so a test for them is an STR test. There are a number of locations where they occur on the Y-chromosome. The test I am using for the Warburton project tests 43 of them. They are known as markers, hence the term 43 marker test. Some tests use as little as 10 markers, though they have limited use.

A mutation is a change in the number of times a short string of DNA is repeated at a particular location on the Y-chromosome. So, for example, the only difference between my profile and that of my genetic cousin Clive is that at location #DYS458 I have 16 repeats and he has 17.

Just how often do these mutations occur? Views on this are still changing. When I started my project the often quoted figure was once in every 500 transmissions from father to son, or a rate of 0.2%. When I look more closely at mutation rates I find a number of sources that don't entirely agree but follow a pattern. Mutation rates would seem to vary considerably between markers, and maybe for some markers, between families. The Warburton DNA Project includes 43 marker tests performed by **DNA Heritage**, and 37 marker tests performed by **Family Tree DNA**. Only 32 markers are common to both tests. Based on the available mutation rate information I am using a rate of 0.23% per marker for tests carried out by **DNA Heritage,** 0.42% for tests at **Family Tree DNA**, and 0.28% for the 32 markers that are common to both tests. This disparity seems quite surprising but the 5 unique **Family Tree DNA** markers are amongst the most volatile.

For each marker there is also documentation on the frequency each possible result occurs. Each possible value is known as an allele, and the proportion of the total results which each allele is found to have is the allele distribution. I have put a **Mutations Table** on my website with the details and sources for mutation rates and allele distributions.

The mutation rate is used to calculate the probability of a common ancestor within so many generations. For those of you with no background in statistics here is a (hopefully) simple explanation. Suppose I have a bag containing 500 balls, of which 50 are blue and the rest red. I now select 10 balls at random. What are the chances of me getting 1 blue ball? If I kept repeating the test I should get 1 blue ball more frequently than any other result, but on occasions I will get no blue balls, and on other occasions two, three, or rarely even more blue balls. By repeating the test many times the frequency of each result can be determined and plotted on a graph. This will tell me what are the chances (the probability) of any given test returning just one blue ball, or any other number of blue balls. Of course there are mathematical formulae to work out the results without having to keep testing it.

In the same way it is possible to calculate the probability for the number of generations back to a common ancestor given the number of mutations between 2 results. Most Recent Common Ancestor (MRCA) calculators have been developed to make this easy, and are available on the web. The one I use most is at www.moseswalker.com. It needs just 3 pieces of information, the number of markers being compared, the average mutation rate, and the number of matches.

A mutation rate 0.3% is one in every 333 transmissions of a marker from father to son. Therefore over 33 markers a mutation should occur, on average, once in every 10 transmissions of the Y-chromosome from father to son. This is one mutation in 10 generations, or if a man had 10 sons, one son would have a mutation at one location. These figures are in fact quite close to the 0.28% mutation rate I am using for the 32 markers that are common between the **DNA Heritage** and **Family Tree DNA** tests used in the Warburton project.

Surnames were introduced around 12-1300 AD, when feudal estates needed them for record keeping. In fact it is almost exactly 750 years since Sir Piers de Dutton built the manor house at Werberton and began to style himself de Werberton, The number of generations this represents depends on the length of a generation. I have seen references to a generation being 20-25 years but these are focussed on inheritance and the first born. In my own line the average of the last ten generations is over 37 years, with one of my ancestors being born when his father was 59 years old. There is probably considerable variation from line to line but I will use an average of 30 years per generation, which gives 25 generations over 750 years.

When calculating the generations to a common ancestor from two matched participants a generation involves two transmissions. A grandson is two generations from his grandfather, but there are four transmission events between two cousins who share a grandfather. Two modern day descendants of Sir Piers de Werberton would be 50 transmissions apart so we might expect to see 5 mutations between them.

Most results are either clear matches, with 4 or fewer markers having a different value from a previous result, or clear mismatches with 9 or more differences from each previous result. A couple of results have fallen in between, giving a below 10% probability of a common ancestor in 25 generations. However this may be a considerably higher probability than the probability of a random match.

The probability of 2 random individuals having the same values for a set of markers can be calculated using the allele distributions. For example if the probability of a specific value at one marker is 90% and the possibility of another specific value at another marker is 80% then the probability that 2 people have the same values is 72% (0.9 x 0.8 expressed as a percentage). If you multiply the probabilities of having the most common allele for each of the 32 common markers, then the probability of 2 random individuals having a complete match is just 6 one thousandths of one percent. The chances of sharing some of the less common alleles are even less, particularly when there are more than 2 participants in the potential group.

We must also acknowledge the possibility that a common ancestor could be a pre-Warburton, or a non-Warburton.The village of Warburton was founded in the 10th century as a fortified settlement near a ford on the river Mersey. I would imagine it was established with a handful of families, and 3-400 years later when surnames were being adopted it would still be largely populated by descendants of these families. This would be particularly true of the males, although wives might come from neighbouring villages. It is possible, therefore, that if 2 sons of a village where many people were already related, should independently adopt the Warburton name they might already have a common ancestor. We therefore have the possibility of a pre-Warburton common ancestor up to 1050 years, or 35 generations ago. As an aside it is worth commenting that while the lords of the manor adopted the Warbruton name when they moved to the village, as it identified them in regional or national circles, an adopter from the lower classes would probably only take the name when he moved away.

The other possibility is a non-Warburton common ancestor. I have noticed it is not unusual for there to be multiple marriages between two families. It must be possible, therefore, that 2 Warburton girls, maybe from different generations, have illegitimate sons by members of the same local family, thus giving them a common non-Warburton ancestor.

The actual probabilities for generations to the most recent common ancestor, based on the appropriate mutation rates for each possible comparison (i.e.between 2 **DNA Heritage** results, 2 **Family Tree DNA** results, or the common marker set) are given in the **Mutations Table**.

The number of repeats can change up or down, and occasionally by more than 1. It could be that due to random changes cancelling each other out, two people who are unrelated finish up with the same profile. We have seen the probability of a random match is low, but in a large population some random matches are inevitable. Therefore matches are only considered meaningful when there is additional information to link two people. A shared surname is such a piece of additional information. This is why most STR studies are surname studies, though there are some locational ones.

A match between a Warburton and a non-Warburton, could be evidence of an illegitimate birth where the father was a Warburton, but additional evidence would be needed to confirm a link.

Because the range of possible generations to a common ancestor is quite wide, it is important to factor in knowledge from genealogical studies wherever possible. For example, if we calculate for myself and Clive with 42 matches out of 43, we can see that our common ancestor is most likely between 5 and 14 generations back , with the most probable result being 9 generations. In fact, based on genealogy, the only possible candidate within the last 10 generations was 9 generations ago. However it must be remembered that there is still a 25% chance he lived over 14 generations ago. See **My Genetic Links** for more about my search for links to my genetic cousins.

The DNA test itself is actually very simple. The hardest part is parting with the money, though there is a discount when it is taken as part of the Warburton Surname Project, and from time to time there are additional promotions. The participant is sent a couple of cotton buds in the post. He wipes these around inside his mouth for 30 seconds, puts them in the container provided, and posts them back. Results are returned within a few weeks.

Although a male Warburton is needed to take the test there are many instances where the genealogist is a female relative. Provided the male is willing to provide his DNA there is no reason why the female genealogist cannot handle all other aspects of participation on his behalf.

Some people might be fearful of the idea of giving their DNA due to unforeseen consequences. Most concerns are groundless. Firstly test samples are normally kept by DNA Ancestry for 3 years in case further tests (e.g. a SNIP test) are requested, but will be destroyed earlier if requested, so that further, unauthorised testing is not possible. The test itself is far too limited for forensic purposes. It is designed to match people, not uniquely differentiate them. Also no medical information can be determined. Remember these tests target the useless junk DNA where changes have no bearing on the persons life, health, or ability to pass on the changed DNA.

However, there is one issue that participants should be aware of. Whilst a test is not detailed enough to prove paternity, it can prove two people are unrelated, which could be a problem if they thought they were. To mitigate this possibility two close relatives should not both participate. The results would not be particularly useful to the project anyway.

Part 4: Using the Results

DNA testing is best viewed as an additional tool in traditional genealogy. We are trying to add information to help us understand our past better. In the Warburton Surname Study I am using traditional genealogy to map extended Warburton families, or clans, and using DNA to group these into genetic clans wherever possible.

A result consists of a string of numbers, one for each marker, and a prediction of the haplotype that would result from a Y-chromosome SNP test. Each of the markers has a name, such as DYS19. The associated number is typically between 10 and 30, and is the number of times a small sequence of DNA is repeated at the location of that marker.

The first step is to group people into genetic clans where all of the results are a match. Whenever two or more results match we have identified a triangle in which the participants' common ancestor is at the apex, the participants form the base, and all the other descendants of the common ancestor are contained within it. To be included in a genetic clan a person should have no more than 5 mismatches from another clan member. More than that would raise uncertainties, though these might be resolved by more results that provided further links.

With a lot of results the pattern of the clans would be instructive for Warburton history. For example if 50% or more of us fell into a single clan, with everyone else falling into relatively small clans, this would suggest that a single person adopted the name originally, and most of us are descended from him. If we were to find several large clans, it would suggest a number of men adopted the name around the same time.

Not all Warburton DNA profiles will have been associated with the surname for 700 years. Whereas DNA profiles always pass from father to son this isn't always the case with surnames. Some clans, particularly smaller ones, will be the result of 'non-paternal events'. These are occasions when a male receives the Warburton name from someone other than his biological father.The rate of such events is apparently about 2% per generation.

There are a number of reasons why a son might not take his natural father's name. Infidelity and illegitimacy are obvious reasons, but it isn't unheard of for a family to take the wife's name if she brings a considerable inheritance to the family. For example the Egerton family who inherited Arley Hall in 1813 from Sir Peter Warburton through his niece, changed their name to Egerton Warburton.

Also a name might simply be adopted, for example because it is a step-father's name, or it is politically expedient. There is a case documented in the London Gazette of 1792 of a Charles Terence Mongon adopting the name Warburton (the name of his maternal cousins) apparently because it would aid his preferment in the protestant Church of Ireland.

Of course the original adoption of the name was a 'non-paternal event'. So except where we have a documented history like in the case of the Warburtons of Arley, we can never be sure of the exact details. An early illegitimacy could start a clan as big as one resulting from an original adoption of the name by a feudal serf.

Once we have identified one or more genetic clans we can explore the clans themselves. Individual clan members are all distant cousins sharing a common ancestor so they will want to find where the link might be. The degree of relatedness will indicate the possibility of finding the link. If the probability of a link since around 1600 AD is high then it is worth looking for the link in parish and other records. Even if the link is earlier it may be possible to show a relationship between two previously unrelated branches.

As an example I indicated earlier that there is a 73% possibility that the common ancestor of my genetic cousin Clive and I is within 10 generations, or post 1600. **My Genetic Links** describes my search for links to Clive and my other genetic cousins.

By looking for earliest known ancestors we may find a clan originates in a particular location. For example Cheshire Warburtons may be distinct from Lancashire ones. Such information may be particularly appealing to overseas Warburtons who want to know more precisely where they originate from.

As well as directing our more traditional genealogical research, the study of genetic clans can employ some new techniques. When there are a few matching results they can be linked together in a phylogenetic tree, which is a simplified family tree in which the location of changes is deduced as far as possible.

As an example, the one difference between my genetic cousin Clive and myself is that he has 17 repeats at marker DYS458 and I have 16. Subsequent results have also had a value of 17, indicating that 17 is probably the original value and the mutation to 16 took place somewhere in my line.

If we tested a new participant with whom I shared a more recent common ancestor, we could pinpoint where the change occurred more accurately, based on whether he had a value of 16, or 17. Furthermore, if the new participant had a different change from me, any future clan member showing the same change would be closely related to him. See the example genetic family tree below.

```
Common Ancestor X
      |     =17     |
      |             |
      |             |
      |    Common Ancestor Y
      |            |     =17      |
      |            |              |
      |           # 17->16     @   new change
      |            |              |
   Clive          Me        New Participant
    =17          =16      =17 + new change
```

=17 or =16   the value of DYS458
#   position of change from 17 to 16 (roughly: it could be anywhere between common ancestor Y and Me)
@ rough position of new change

As more and more participants are shown to be clan members the genetic family tree will become richer. Combine this with knowledge of the genealogical family tree and you can see how, over time a DNA test will provide more and more information about where a new participant fits in. It is my hope that I might reach this position over the next few years.

In the Warburton One-Name Study I am using traditional genealogical research to build a series of family trees. These family tree are the inverse of a traditional family tree in that they follow multiple lines of descent from a single common ancestor. As such they mirror the triangles produced from using DNA profiles to identify a genetic clan so they are published as Warburton clans. A DNA triangulation (genetic clan) might encompass part of a Warburton clan, a whole clan, or even several clans. The Warburton clans might also help identify DNA test subjects.

My objective is to assign a genetic profile to each of the Warburton clans, that I document. A single Y-chromosome DNA result will link a DNA profile to a Warburton clan, but it will give no clue as to how long the profile and the clan have been linked. Clues may come from genealogy. Many illegitimacies are recorded and so would give a probable time of linking. However the only way to get any clarity is to find a matching profile, and determine the common ancestor, or at least when he might have lived. It is not always possible to be precise, but boundaries can be determined.

When the match is with a member of a different clan then the two clans are linked. The common ancestor may be within the clan with the deepest known history, or he might pre-date the earliest known ancestor in each clan. A judgement may be made based on the closeness of the match. Once a link of this kind is made there is little point in another member of the clan being tested as he should match the existing results. There might be a small benefit from refining the phylogenetic tree, but this could be offset by the risk of uncovering an unknown, recent, non-paternal event.

When a match is between two members of the same Warburton clan then the profile is determined for the descendants of their common ancestor, but they could still be part of a segment of the clan which is defined by a non-paternal event.

Whenever an unmatched profile is obtained my objective is to achieve a match to enable triangulation. One strategy is to hope a match turns up. However, given the current rate of testing this might take a while. The alternative is to look where a match should be. If the unmatched profile is linked to a family tree of sufficient breadth it may be possible to identify a distant cousin who should be a match, and to test him. Focussed testing like this might produce quicker results. though it is possible that the new test is not a match if there has been a recent 'non-paternal event'. It might take a third or fourth test to isolate the problem. When a match is found it might not be to the original participant, who would now know his link to the clan is through a non-paternal event.

Another line of research is to look for matches beyond the Warburton surname. There are websites where profiles can be stored and matched. It is possible a match may be close enough to determine the surname of the father in a non-paternal event. However it is also possible that such a match is merely random so other evidence is needed to prove a match.

Y-chromosome DNA profiles can be extremely useful in the study of family history if they are used in addition to the traditional tools of genealogical research. My objectives in the Warburton One-Name Study is to continue to exploit this synergy to develop an ever broader understanding of the various Warburton clans and what links may exist between them.

References

My own interest in this area grew out of an interest in Ancient History. I began with Ancient Egypt and then began reading about earlier subjects, including evolution and the history of climate. One book I came across was **Out of Eden** by Stephen Oppenhiemer. The subject of the book was what I now know to be a new 'science' called phylogeography. This combines phylogenetics with traditional archaeology to study the ancient migrations of peoples. The startling conclusion of the book was that all non-Africans in the world are descended from a small group of humans that left Africa 80,000 years ago.

I then came across two books by Professor Bryan Sykes, **The Seven Daughters of Eve**, and **Adam's Curse**. These books have very readable discussions of the science, and lots of interesting anecdotes. For example he shows how Thor Hyerdahl's Kon Tiki expedition to show how Polynesia could have been populated from South America was a waste of time because genetics proves that the Polynesians came from China, probably via New Guinea. However, I found the pseudo-life descriptions of the seven European clan mothers a bit contrived.

If you are interested in the genetics and evolution, I also enjoyed **The Selfish Gene** by Richard Dawkins.

Professor Bryan Sykes has set up a testing company called **Oxford Ancestors** (www.oxfordancestors.com/) to allow people to get their DNA tested, so I got a test. My mitochondria test was interesting. It is discussed in the **Results** chapter. However, my Y-chromosome test turned out to be an STR test on a limited number of bases. Nevertheless my deep ancestry was deduced from this. I seem to match the most common Western European type, known as the Atlantic Modal Haplogroup (AMH).

I then read a book called **DNA and Family History** by Chris Pomery and realised my test result is not terribly useful for genealogy. This is a very useful book and has an associated website (www.dnaandfamilyhistory.com. To make it easy to look at other surname DNA projects I have included my site in a webring of similar sites. In particular I recommend the Davenport website (www.davenportdna.com/). Davenports have similarities with Warburtons, with similar numbers, a Cheshire origin, and a story of Norman ancestry. The International Society of Genetic Genealogy (ISOGG) also have an interesting site (www.isogg.com/).

Recently both Sykes and Oppenheimer had produced books on the origins of the British peoples. Sykes' 'Blood of the Isles' is perhaps the more readable, but Oppenheimer's 'The Origins of the British' is the more detailed work and I have used it as the basis for my comments on the haplotype predictions for Warburton Surname DNA Project participants.