

Ludlow Genealogy

www.ludlowfamilytree.org

DNA Studies

Report by Tony Ludlow

E-mail: tony@ludlowfamilytree.org

11th February 2007

1 Introduction

oxfordancestors.com carry out tests of the Y-chromosome because this passes from father to son, like the surname. To remind you of the genetics, males have one Y and one X-chromosome; females have two X chromosomes but no Y. So the Y-chromosome in any male must have come from his father.

The analysis focuses on the Y-Line which consists of 10 sites (or loci) along the chromosome. At these loci the DNA consists of multiple copies of junk DNA, i.e. sequences that do nothing. Because they do nothing, there are no medical implications and there is no pressure to speed up or slow down changes (or mutations). As a result, the mutations that occur are just copying errors which seem to be random and, on average, there is one mutation in each locus every 500 generations so, over all 10 loci, there is 1 mutation in every 50 generations.

At locus A, there are from 12 to 19 copies of junk DNA, at B there are 10 to 17 and so on. Brothers should have identical scores at each of the ten loci but, over the generations, mutations will arise in their descendants. Because of the slow rate of mutations, we should expect one or two differences to arise in 600 years. Any more than three differences would imply that the individuals were not related.

2 A Possible DNA family tree

In the analyses, so far, there are conflicting pointers. There are rather more mutations than we might expect, but a 'DNA tree' can be constructed showing a 'main line' from which other Ludlows differ by only one or two mutations (Fig. 1). This DNA tree fits with other evidence.

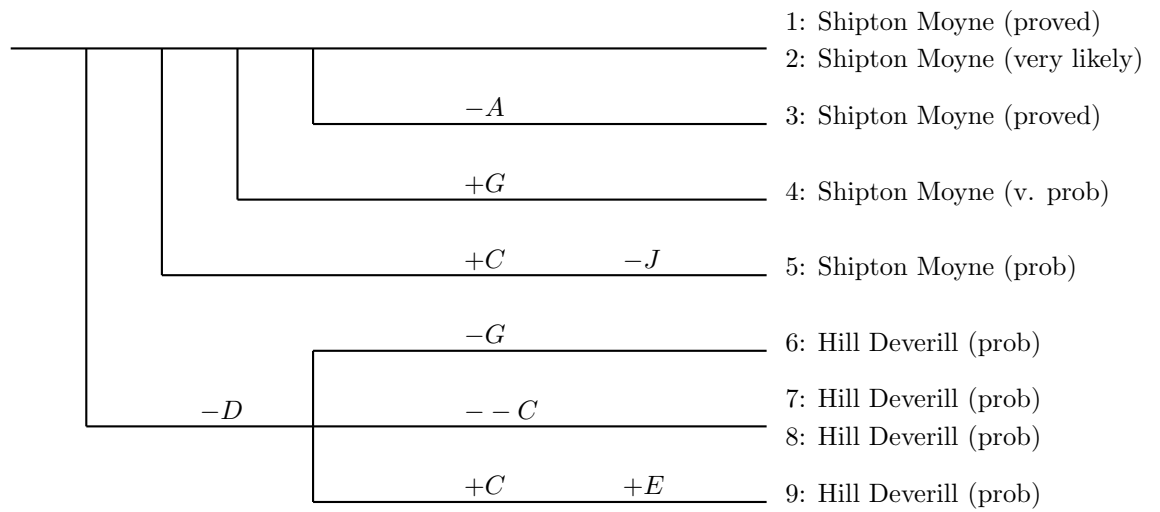


Figure 1: A possible family tree based on DNA evidence

Figure 1 shows one of many possible DNA trees that could explain the results, but it is the simplest. If the tree is true, then the following conclusions can be drawn.

The top line represents the original Ludlow DNA which survives unchanged in two individuals (1 and 2). Individual 3 differs because there has been one copying error at locus A and the mutation lost one copy of junk DNA.

We know from normal genealogical work that individual number 1 in Fig. 1 is related to individual 3 and that they share a common Great Great Great Grandfather, so their closest common paternal ancestor (CPA) was 5 generations back (b. 1723) and the $-A$ mutation must have occurred in those 5 generations (280 years)

Individual 4 seems to have gained one copy at locus G.

Individual 5 seems to have gained a copy at locus C and lost one at locus J.

Individuals 6–9 all seem to be related because they all have lost one copy of the DNA at locus D. Since the mutation applies to all of them it is logical to conclude that they share a common parental ancestor (I'll call him CPA–D) who had already had the mutation at locus D. CPA–D, in turn had an earlier common ancestor linking him to the original family (on the top line).

The four descendants of CPA–D each carry a further mutation. One of the ancestors of individual 6 has lost a copy at locus G. Individuals 7 and 8 are known to be second cousins, so their most recent CPA was born after the loss at locus C. The double ‘–’ sign indicates that two copies of C were lost. It is quite common for two copies to be gained or lost as a single mutation, but a difference of three copies almost always means two mutations.

Finally, the ancestors of individual 9 have experienced two mutations: gaining a copy at locus C and a copy at locus E.

One interesting thing about this tree is that it fits what we know of the three big Ludlow families: Shrewsbury, Hill Deverill and Shipton Moyne. Individuals 1 to 5 are believed to be descendants of the Shipton Moyne family. We can be sure in the case of 1 and 3 because their relationship has been proved at the Royal College of Arms. Individual 2 is clearly related to 1, with identical scores at each locus. Individual 4 differs from 1 and 2 at only one locus. There is no reason to doubt that these five individuals are closely related to 1 who, we know is from Shipton Moyne. Since the Shipton Moyne family was founded a little before 1545, there has been up to 460 years for these mutations to happen.

Individual 5 could well be related to individuals 1–4 because the probability of having two or more mutations in 460 years (about 15 generations) is about 0.10 or 10%. We may conclude that 1–5 are from the same family which we know to be the Shipton Moyne Ludlows.

Individuals 6–9 form a group because they share the $-D$ mutation, which could be an ancient mutation. Individuals 7 and 8 come from Ireland where several Hill Deverill families have been known to settle and 9 comes from the USA where, again, there were several branches of the Hill Deverill family. Again, the probability of two or more mutations, following $-D$ is about 0.10.

Individual 6 has proved an enigma. The normal genealogical research has drawn a blank although Keith Ludlow thought he had evidence that individual

6 was from Hill Deverill. That would agree with the DNA evidence but is not enough to corroborate it. It seems fairly clear, though, that individual 6 is not a Shipton Moyne Ludlow.

3 Alternative DNA trees

Section 2 is headed “A possible DNA tree” because many alternatives can be constructed. For example, we could swap the branching point for individual 3 with that for individual 4. The only evidence for the choice shown in Fig. 1 is that we know the $-A$ mutation was recent (after 1723) but we don’t know when the $+G$ mutation happened. Similarly, we don’t know the order of the $+C$ and $-J$ mutations for individual number 5. Since there are two mutations on this path, it is logical to allow a longer time (shown in Fig. 1 by an earlier branch point) but we cannot be sure.

Another change that cannot be ruled out is the difference between the Hill Deverill and Shipton Moyne families. I have drawn it so that the Hill Deverills had a mutation in which they lost a copy at locus D but it could equally well be that the Shipton Moynes gained a copy at locus D. My reason for drawing it the first way round is that there were two individuals (1 and 2) who were not closely related on the normal genealogical evidence who had identical DNA scores. That implies that they have a quite distant CPA (Common Paternal Ancestor) and so may represent the ancient DNA sequence. Two individuals from Hill Deverill (7 and 8) also had identical DNA scores but were second cousins so the mutations could have occurred quite recently.

However, it wouldn’t have made the slightest difference to our conclusions: that there are two groups and that these match the Hill Deverill and Shipton Moyne split.

This tree has the merit that it requires the minimum number of independent mutations. The same mutation ($+C$) occurs twice (for individuals 5 and 9). But that is the only duplication. For many other arrangements, we would have to postulate that mutations such as $-D$ had occurred more than once.

It turns out that there is a straightforward way of finding the best tree, discussed in the next section.

4 Finding the best tree

The method used here is to compare each Y-Line with every other. This is done with a Prolog program and the output of the program is shown below.

Beginning with individual 1 as the base for comparisons, we compare his Y-Line scores with those of individual 2, then 3 and so on. The differences in number of copies are shown for each locus under the heading “Differences”. The sum of the differences are shown next. An alternative measure is the number of loci affected. This is shown under the heading “No. of Loci”. Finally, the squares of Diffs and Loci are calculated, followed by the sum of their squares.

Since individuals 1 and 2 have identical Y-Lines, the second block is identical to the first.

The third block of Differences shows the effect of using individual 3 as the base for comparison. The most noticeable difference is that the total sums of squares has risen from 37 with individual 1 to 72 when individual 3 is used as the base for comparisons.

The sum of squares is used as a measure of variation in many contexts and the best tree is the one with the lowest sum of squares, as shown in Fig. 1.

One reason the sum of squares is lower when number 1 is used as the base for comparisons might be that number 1 and 2 are identical. To test this, I deleted number 2 and re-ran the program. The effect on sums of squares was negligible, with the highest sum of squares dropping from 101 to 97.

Once we know which block has the lowest sum of squares it is a simple matter to construct the tree. You draw a line across the top of the figure to represent the 'original Y-Line'. Then, using the column headed Sum of Diffs, you identify the rows with the lowest differences and connect them to the top line, giving an early branch point if there are two or more mutations from the original Y-Line, and a later branch point if there is only one.

An important question is: "How many mutations should we expect in the time since the Hill Deverill and Shipton Moyne families divided?". There are various checks that can be made and these are examined in section 6

5 Output from analysis program

THE ANALYSIS OF DIFFERENCES BETWEEN Y-LINES OF 9 INDIVIDUALS
(OUTPUT FROM PROLOG PROGRAM)

```

Sums of differences and number of loci affected
                Sum of Sum of  Sq of  Sq of
Differences          Diffs  Loci  Diffs  Loci

A B C D E F G H I J

[0,0,0,0,0,0,0,0,0,0]      0      0      0      0  for 1 v 2
[1,0,0,0,0,0,0,0,0,0]      1      1      1      1  for 1 v 3
[0,0,0,0,0,0,0,1,0,0]      1      1      1      1  for 1 v 4
[0,0,1,0,0,0,0,0,0,1]      2      2      4      4  for 1 v 5
[0,0,0,1,0,0,1,0,0,0]      2      2      4      4  for 1 v 6
[0,0,1,1,1,0,0,0,0,0]      3      3      9      9  for 1 v 7
[0,0,2,1,0,0,0,0,0,0]      3      2      9      4  for 1 v 8
[0,0,2,1,0,0,0,0,0,0]      3      2      9      4  for 1 v 9
Total sums of squares:                37      27

```

[0,0,0,0,0,0,0,0,0,0]	0	0	0	0	for 2 v 1
[1,0,0,0,0,0,0,0,0,0]	1	1	1	1	for 2 v 3
[0,0,0,0,0,0,1,0,0,0]	1	1	1	1	for 2 v 4
[0,0,1,0,0,0,0,0,0,1]	2	2	4	4	for 2 v 5
[0,0,0,1,0,0,1,0,0,0]	2	2	4	4	for 2 v 6
[0,0,1,1,1,0,0,0,0,0]	3	3	9	9	for 2 v 7
[0,0,2,1,0,0,0,0,0,0]	3	2	9	4	for 2 v 8
[0,0,2,1,0,0,0,0,0,0]	3	2	9	4	for 2 v 9
Total sums of squares:			37	27	

[1,0,0,0,0,0,0,0,0,0]	1	1	1	1	for 3 v 1
[1,0,0,0,0,0,0,0,0,0]	1	1	1	1	for 3 v 2
[1,0,0,0,0,0,1,0,0,0]	2	2	4	4	for 3 v 4
[1,0,1,0,0,0,0,0,0,1]	3	3	9	9	for 3 v 5
[1,0,0,1,0,0,1,0,0,0]	3	3	9	9	for 3 v 6
[1,0,1,1,1,0,0,0,0,0]	4	4	16	16	for 3 v 7
[1,0,2,1,0,0,0,0,0,0]	4	3	16	9	for 3 v 8
[1,0,2,1,0,0,0,0,0,0]	4	3	16	9	for 3 v 9
Total sums of squares:			72	58	

[0,0,0,0,0,0,1,0,0,0]	1	1	1	1	for 4 v 1
[0,0,0,0,0,0,1,0,0,0]	1	1	1	1	for 4 v 2
[1,0,0,0,0,0,1,0,0,0]	2	2	4	4	for 4 v 3
[0,0,1,0,0,0,1,0,0,1]	3	3	9	9	for 4 v 5
[0,0,0,1,0,0,2,0,0,0]	3	2	9	4	for 4 v 6
[0,0,1,1,1,0,1,0,0,0]	4	4	16	16	for 4 v 7
[0,0,2,1,0,0,1,0,0,0]	4	3	16	9	for 4 v 8
[0,0,2,1,0,0,1,0,0,0]	4	3	16	9	for 4 v 9
Total sums of squares:			72	53	

[0,0,1,0,0,0,0,0,0,1]	2	2	4	4	for 5 v 1
[0,0,1,0,0,0,0,0,0,1]	2	2	4	4	for 5 v 2
[1,0,1,0,0,0,0,0,0,1]	3	3	9	9	for 5 v 3
[0,0,1,0,0,0,1,0,0,1]	3	3	9	9	for 5 v 4
[0,0,1,1,0,0,1,0,0,1]	4	4	16	16	for 5 v 6
[0,0,0,1,1,0,0,0,0,1]	3	3	9	9	for 5 v 7
[0,0,3,1,0,0,0,0,0,1]	5	3	25	9	for 5 v 8

[0,0,3,1,0,0,0,0,0,1]	5	3	25	9	for 5 v 9
Total sums of squares:			101	69	

[0,0,0,1,0,0,1,0,0,0]	2	2	4	4	for 6 v 1
[0,0,0,1,0,0,1,0,0,0]	2	2	4	4	for 6 v 2
[1,0,0,1,0,0,1,0,0,0]	3	3	9	9	for 6 v 3
[0,0,0,1,0,0,2,0,0,0]	3	2	9	4	for 6 v 4
[0,0,1,1,0,0,1,0,0,1]	4	4	16	16	for 6 v 5

[0,0,1,0,1,0,1,0,0,0]	3	3	9	9	for 6 v 7
[0,0,2,0,0,0,1,0,0,0]	3	2	9	4	for 6 v 8
[0,0,2,0,0,0,1,0,0,0]	3	2	9	4	for 6 v 9
Total sums of squares:			69	54	

[0,0,1,1,1,0,0,0,0,0]	3	3	9	9	for 7 v 1
[0,0,1,1,1,0,0,0,0,0]	3	3	9	9	for 7 v 2
[1,0,1,1,1,0,0,0,0,0]	4	4	16	16	for 7 v 3
[0,0,1,1,1,0,1,0,0,0]	4	4	16	16	for 7 v 4
[0,0,0,1,1,0,0,0,0,1]	3	3	9	9	for 7 v 5
[0,0,1,0,1,0,1,0,0,0]	3	3	9	9	for 7 v 6

[0,0,3,0,1,0,0,0,0,0]	4	2	16	4	for 7 v 8
[0,0,3,0,1,0,0,0,0,0]	4	2	16	4	for 7 v 9
Total sums of squares:			100	76	

ary

[0,0,2,1,0,0,0,0,0,0]	3	2	9	4	for 8 v 1
[0,0,2,1,0,0,0,0,0,0]	3	2	9	4	for 8 v 2
[1,0,2,1,0,0,0,0,0,0]	4	3	16	9	for 8 v 3
[0,0,2,1,0,0,1,0,0,0ary]	4	3	16	9	for 8 v 4
[0,0,3,1,0,0,0,0,0,1]	5	3	25	9	for 8 v 5
[0,0,2,0,0,0,1,0,0,0]	3	2	9	4	for 8 v 6
[0,0,3,0,1,0,0,0,0,0]	4	2	16	4	for 8 v 7

[0,0,0,0,0,0,0,0,0,0]	0	0	0	0	for 8 v 9
Total sums of squares:			100	43	

[0,0,2,1,0,0,0,0,0,0]	3	2	9	4	for 9 v 1
[0,0,2,1,0,0,0,0,0,0]	3	2	9	4	for 9 v 2
[1,0,2,1,0,0,0,0,0,0]	4	3	16	9	for 9 v 3
[0,0,2,1,0,0,1,0,0,0]	4	3	16	9	for 9 v 4
[0,0,3,1,0,0,0,0,0,1]	5	3	25	9	for 9 v 5
[0,0,2,0,0,0,1,0,0,0]	3	2	9	4	for 9 v 6

[0,0,3,0,1,0,0,0,0,0]	4	2	16	4	for 9 v 7
[0,0,0,0,0,0,0,0,0,0]	0	0	0	0	for 9 v 8
Total sums of squares:			100	43	

6 What is the probability of these individuals being related?

An important question is: “How many mutations should we expect in the time since the Hill Deverill and Shipton Moyne families divided?”.

Rare events, such as Y-Line mutations, follow what is called a Poisson distribution. There are independent statistical checks, based on the normal mutation rate and these can be used to see if the conclusions are plausible.

We know from other studies that we would expect an average, over all 10 loci, of 1 mutation in every 50 generations. Are there too many or too few mutations in our study to match this genetic clock?

6.1 Methods of doing the sums

You can skip this section if you don’t like sums.

The Poisson distribution allows us to calculate the probability of observing 1 mutation, over a given period, the probability of observing 2 mutations in that period, and so on.

$$p(X) = \frac{\lambda^X e^{-\lambda}}{X!} \quad (X = 0, 1, 2, \dots) \quad (1)$$

where $p(X)$ is the probability of observing a mutation exactly X times in a given number of generations. λ (pronounced lambda) is the expected rate for the given number of generations G . It can be a bit confusing because, sometimes we talk about the rate per generation and sometimes we talk about the number of mutations in G generations. λ is the second of these two eg. the expected number of mutations per G generations.

For any particular case we can calculate λ from:

$$\lambda = \eta GL \quad (2)$$

where η (pronounced eta) is the rate of mutations per generation; G is the number of generations and L is the number of ‘Lines’ being compared (see below).

Again we have an opportunity for confusion. The average mutation rate over the ten loci is one mutation in 50 generations, so

$$\eta = \frac{1}{50} = 0.02 \quad (3)$$

but we will sometimes consider the number of generations from when the Hill Deverill family started (c. 1399) to now (about 25 generations), and we shall sometimes calculate the number of generations *between* Robin Ludlow and me (about 50 generations because we have to count 25 generations for the ‘Line’ from Robin back to Chipping Camden and 25 generations for the ‘Line’ from Chipping Camden down to me).

To make it easier to understand I shall introduce another term: L is the number of ‘Lines’ being considered. When we are considering the time it took for one mutation to occur in one line of ancestors we set $L = 1$. When comparing Robin and me, we set $L = 2$. And when we are comparing 9 lines of ancestors, we set $L = 9$

And I shall be very careful to make clear which value of L we need. But be careful when using the sheets that Oxford Ancestors send out with the results. On the sheet headed the “The Poisson distribution” they give a useful graph which is equivalent to setting $L = 2$. In the main leaflet, they give an example of Jim and Bob which is also equivalent to setting $L = 2$. In the second example (the Dyson family) they are, in effect, using $L = 1$. What they don’t tell you is that they also change the rate η in between examples, so that they are using $\eta = 0.04$ for the graph.

Putting all this together, we have:

$$p(X) = \frac{(\eta GL)^X e^{-(\eta GL)}}{X!} \quad (X = 0, 1, 2, \dots) \quad (4)$$

You can do these calculations on a spreadsheet.

On an Excel spreadsheet you need to:

1. Enter cells in column A as follows:
 - (a) Enter A5 as **Generations**
 - (b) Enter A6 as **Rate per generation**
 - (c) Enter A7 as **Number of Lines**
 - (d) Fill in a number of cells, say A8, . . . A20 with values for the Number of mutations (X above), i.e. 0, 1, . . . 12
2. Enter cells in column B as follows:
 - (a) Set the whole row B5, . . . P5 with the number of generations, i.e. 1, 5, 10, 15, . . . 70
 - (b) Set B6 with the value of 0.02, the mutation rate per generation (see η above)
 - (c) Set B7 with the Number of Lines (see L above)
 - (d) Set B8 with =POISSON(\$A8, (\$B\$5*\$B\$6*\$B\$7), FALSE)
 - (e) Copy B8 down to the rectangular range: B8, . . . P20

This will create a table showing you the probabilities for each combination of number of mutations (rows) and number of generations (columns). Don't change the mutation rate from 0.02, but do remember to change the number of Lines when appropriate.

To complete the story, it may be useful to show some of the properties of the Poisson distribution (Table 1).

Mean	$\mu = \lambda$
Variance	$\sigma^2 = \lambda$
Standard Deviation	$\sigma = \sqrt{\lambda}$
Moment coefficient of skewness	$\alpha_3 = 1/\sqrt{\lambda}$
Moment coefficient of kurtosis	$\alpha_4 = 3 + 1/\lambda$

Table 1: Properties of the Poisson distribution

6.2 Results

William Ludow of Hill Deverill was born around 1399 and he moved to Hill Deverill in 1439. That was about 600 years ago and the Shipton Moyne and Hill Deverill families must have divided by then. We may assume that that is about 25 generations ago, so we need to look at the column headed 25 generations.

Using the Excel table described in the previous section, I was able to answer the following questions.

- What is the probability of individual 3 being related to 1 and 2? We need to use Lines, $L = 2$ and, since there is only one mutation, the answer to the question is that the probability is 0.37, so there is a 37% chance of observing one difference between individuals 1 and 3. The same probability is calculated for the difference between 1 and 4.
- What is the probability of each of the individuals 5, 6, 7, and 8 being related to individual 1 and the answer is 0.18 or 18%.
- What is the probability of individual 9 being related to 1. The answer is 0.06 or 6%

These values are all feasible, but when we come to ask whether 9 is related to 5, the answer is 0.0013 or 0.3%. That is a result which is unlikely to occur within 625 years, but not impossible.

We conclude that there are more mutations than expected, but given the small numbers, it is not conclusive.

What is needed is a global test that takes account of all the individuals at once.

One test, based on the Dyson example in the Oxford Ancestors leaflets, suggests that we should take the total number of mutations from Fig. 1: $N_m = 9$ and the total number of Lines, $L = 9$. Then the estimated time to the closest

CPA is $50N_m/L$ which gives us $50 \times 9/9 = 50$ generations, but I can see no justification for this calculation.

It seems to me a better approach is to argue that comparing 9 Lines is like comparing 2. So, setting Lines, $L = 9$ the column headed 25 generations and row $X = 9$, we see the probability $P = 0.0232$. That means that there is a 1 in 43 chance of getting this result by chance. How should we interpret this?

Scientists usually test the ‘null hypothesis’ which means that they are looking to see if the results they get are simply chance, or whether there is something important going on. They use the 1 in 20 level as a boundary, saying that the results are statistically significant *if they can reject the null hypothesis*.

In the present case, the null hypothesis is that the 9 individuals analysed so far are related to each other in the last 25 generations. What we want to find is that the chance of them being related is more than 1 in 20. The observed value is 1 chance in 43.

There may be several things going on. The closest CPA could be further back than 25 generations. If we chose 31 generations for the null hypothesis, the chances would be 1 in 18. So we should ask if there is a common paternal ancestor 31 or more generations back.

Another possibility is that one or more of the individuals are not related. What would happen if we dropped individual 9? The answer is that X would drop from 9 to 6 and the number of Lines would drop from 9 to 8. The probability at 25 generations is now $P = 0.1042$, over 1 in 10.

But is it really realistic to say that individual 9 is not related? I suspect two things are going on. Firstly, all statistical tests assume that the individuals are chosen at random and are independent of each other. But the nine individuals included in this sample are self-selected, not randomly selected. For example, individuals 7 and 8 are second cousins and came into the analysis because another cousin introduced them to Robin Ludlow. Dropping either of them is enough to retain the null hypothesis, that they are related.

So, we cannot prove that this group of Ludlows are related, but the sample is small and there is no strong evidence that they are not. The numbers involved, and the rarity of mutations, mean that the clock is not accurate enough to be sure.