Surname Simulations, DNA, and Large-Descent Families

by Dr John S Plant (Member 4890)
and Prof Richard E Plant (Member 6100)

Our progress with surname simulations has been outlined in JoONS articles,[1][2][3] on the Guild website[4][5] and elsewhere.[6] In this article, we consider some corroborating Y-DNA evidence and address some difficulties in unraveling descent lines, especially for large families in large surnames.

A Brief Summary of our Simulations

We have developed a computer model for the simulation of English surnames. Large surnames are more regular than are rare surnames in terms of patterns of emigration and population growth overseas and in the UK. Our simulations predict that, for common surnames, the overseas population will generally be around four times that in the UK, with a spread of uncertainty that arises from random chance. The predictions of our simulations are largely borne out by the available documentary and DNA evidence.[5]

Chance is a main factor in determining whether a descent family dies out in the male line, remains small, or grows large.[1][4] By a descent family, we mean the descendants down male lines from a single male progenitor in (say) 14th century England. The largest predicted size for a descent family in the UK can be affected by other factors besides random chance. These include the possibility, for example, of the first men having several successive wives who together produce more than the usual number offspring: this increases the chances that a subsequent descent family will survive and thrive. Also, there is an effect from how much the general population in a region has increased.[2][4]

We have input known historical population data into our model and found, for example, that past conditions favour large descent families in the county of Staffordshire more than for England as a whole. There are also indications of particularly high surname growths in parts of south-east Lancashire and West Yorkshire.[4] Some of this is illustrated in Figure 1 and fuller details are given elsewhere.[4][5]

Our simulations predict that there may be many more origins to a surname than the number whose male lines survive.[1] Note that by a single-origin surname, we mean one whose *living* population had just one origin to the name; there could have been other origins whose descents died out.

The case of a single-origin surname is represented at the extreme right of Figure 1. This corresponds to the label *1* on the x-axis, meaning that there is just one surviving descent family in the surname. The largest predicted size in the UK, under Staffordshire growth conditions, is then represented by the right-most blue triangle (labeled Staffs High). This particular blue triangle has a y-axis value of just over ten thousand people. However, this is an extreme event and a single-origin surname is more likely to be much smaller. For example, in 90 percent of cases, assuming Staffordshire growth conditions, the largest descent family size is predicted to be below 10,000 people (right-most brown triangle, labeled Staffs 90pc).

Rare Surnames

As an example of a rare surname, we can consider one with a UK population of 100 people. This corresponds to the value 100 on the y-axis of Figure 1. Under English growth conditions, such a surname is predicted to contain three or fewer descent families. This limit is represented by the result that, at *3* on the x-axis, the blue square (labeled England Low) has a y-axis value of about 100. This means that, in English growth conditions, three surviving descent families can be expected to have a combined population as low as 100, but this is a statistically-unlikely event involving three unusually small descent families.

Statistically extreme results are omitted from Figure 1 by ignoring the blue squares and blue triangles. The right-most brown diamond (England 10pc) has a value of about 100 on the y-axis. This signifies that only 10 percent of single-origin surnames are predicted to be smaller than this.

Further simulations (not shown here) do not rule out that a surname sized 100 could conceivably, as a rare event, contain more than three descent families. This might arise in a place where lower growth factors pertain, such as in the county of Wiltshire.[4][6]

A possible shortcoming of these simulations is that we are neglecting that a surname might have been adopted, or found its way to England, relatively recently. That would provide an

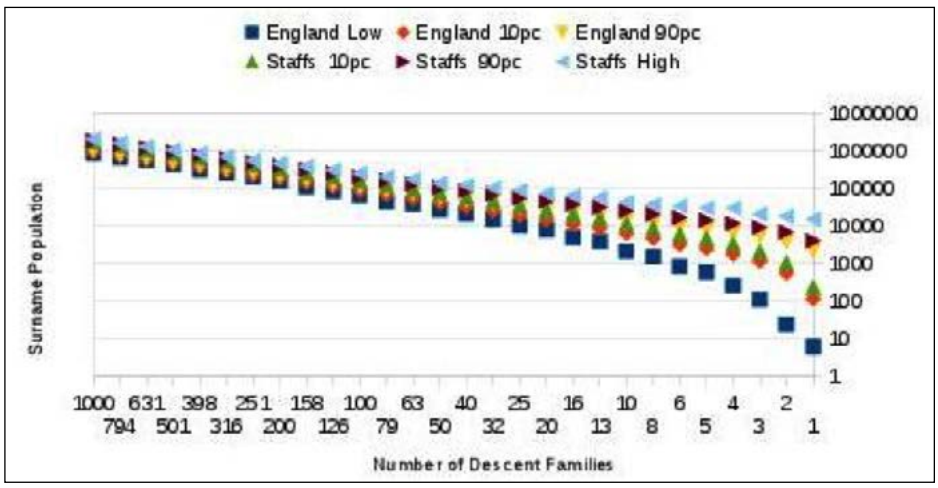

Figure 1: Predicted UK populations of surnames containing various numbers of surviving descent families. As well as upper and lower limits, 10th and 90th percentiles are shown.

alternative explanation for the low population of a surname that is rare in the UK. For a large descent family, we can be more confident that it originated early, allowing more time for it to grow large fortuitously in the UK.

Limits of DNA Corroboration

We can compare the predicted descent family sizes from our simulations with some observed results for real surnames. Estimates of real descent family sizes can be obtained by Y-DNA testing though the process is not without problems. We here consider some data for the Smith surname and illustrate a particular difficulty that arises for estimating large descent family sizes in prolific and very common surnames.

First, it is important to stress that the men selected for this type of DNA test need to be a random sample of the surname. Otherwise, the findings will not typify the surname as a whole. In practice, random samples are rarely obtained though some suitable DNA results are given in a 2008 research paper by King and Jobling (K&J) for variously sized surnames.[7]

In analysing the Y-DNA results, one starts by identifying a descent cluster of closely matching men in a random sample of the surname. This then needs to be augmented by adding others who do not even nearly match — how many more depends on an assumed rate of non-paternity events (NPEs).[5]

K&J used some *ad hoc* rules to assess that there was a "true descent cluster" (*sic*) of nine nearly matching men for Smith. Depending on how they are interpreted, these rules may correspond to as much as 15.5 percent of their random sample which amounts to a UK descent family size of around one hundred thousand people. This very large number exceeds our simulations' predicted upper limit of around 10,000 people in a single family.

However, the rather arbitrary nature of a deduced cluster size is illustrated in Figure 2. This shows a so-called phylogenetic tree which we have obtained by applying MEGA6 software[8] to the raw K&J DNA data for Smith.[7] Near the top of Figure 2, there are three exactly matching men, denoted Sm42, Sm43 and Sm41. Around this core, K&J's *ad hoc* rules deemed that there was an agglomeration of a further six close matches. Such a biological cluster is not readily apparent in Figure 2.

There is also, in fact, some similar clustering in K&J's control sample of 110 men, whom they selected randomly from the general population of England. Their control sample has three adjacent pairs of exactly matching men and these could be taken together as a "split core" for a cluster.[6] This clustering adjoins the Western Atlantic Modal Haplotype (WAMH) and near here, in particular, such clustering is well known *not* to correspond necessarily to a "true descent cluster" within a surname. Indeed, each man in the control sample has a different surname. The cluster that K&J identified for Smith was similar to that in the control sample and also near the WAMH.

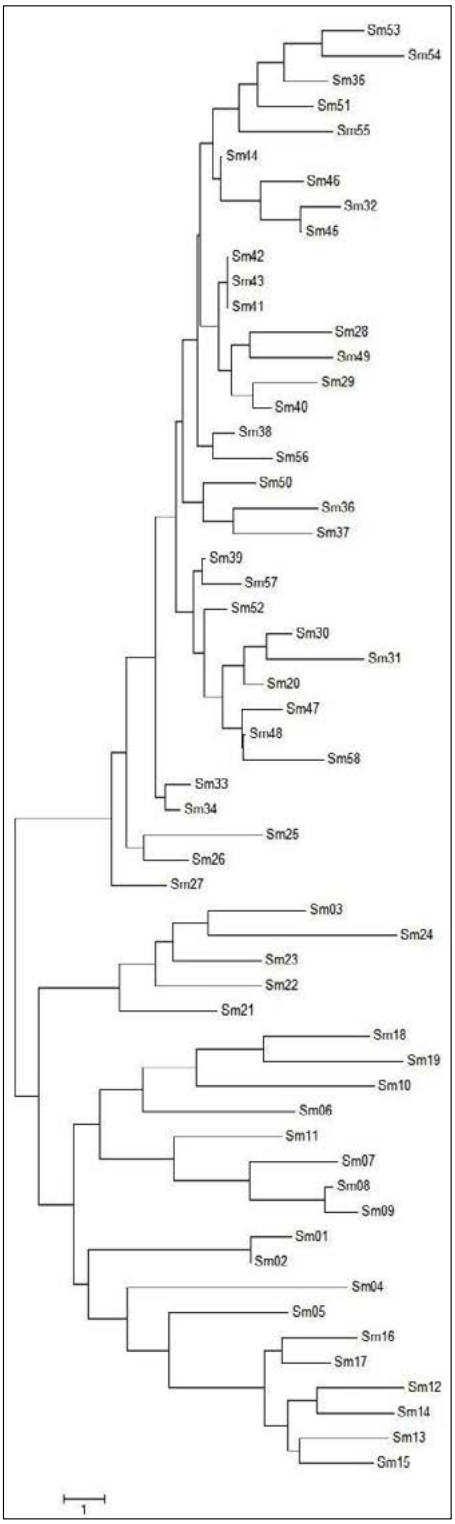

Figure 2: MEGA6 phylogenetic tree of K&J's DNA data for Smith.

K&J were well aware of this phenomenon, but they went on to couch their explanation in technical detail that might well not be meaningful to a non-statistician. Briefly, they computed a set of statistics denoted *h*, F_{ST}, and R_{ST}. Each of these is a measure of the genetic structure (roughly speaking, the amount of relatedness) in each surname. They also plotted analyses of the F_{ST} and R_{ST} statistics and they denoted, in a graph, that the overall genetic structure of the Smith sample was not significantly different from that of their control sample. They found the same lack of statistical significance for other large surnames.

In short, K&J's *ad hoc* rules for identifying "true descent clusters" give a result that apparently discredits our computer simulations in the case of Smith. However, following their further analyses, K&J concluded that their Y-DNA findings provide no statistically significant evidence of UK descent families that are larger than ten thousand people. With this eventual qualification, K&J's findings are consistent with our simulations.

Clustering Rules

K&J's *ad hoc* rules were designed for a case where 17 Y-STR markers are measured for each man, together with a few known Y-SNPs for each. As some readers will already know, in a descending line of fathers and sons, a Y-STR will typically mutate much more often than a Y-SNP. We refer the reader to other texts for a more detailed description of these mutations to the Y-chromosome, which are passed down male lines of descent.[9]

K&J used Y-SNPs to allocate their tested men into some known common Y-DNA haplogroups (*aka* clades). A haplogroup groups together those descended from the same putative forefather, traditionally in prehistoric times. For those men who were found to belong to either one of two specific haplogroups (R1b1 or I), K&J's rules required that there need to be at least three men with an exactly matching set of Y-STRs. Only then would they allow it, in the case of these two particular haplogroups, to qualify as a core for a true cluster.

When more Y-STR markers are measured, such a core of exact matches will typically split apart. For example, for a haphazard sample of men bearing our own common surname Plant, we have a core of *exact* Y-STR matches at the 12 marker level; but, this is resolved into

nearly matching men, when the number of measured Y-STRs is increased from 12 to 37.[10] Elsewhere,[5] we accordingly describe different considerations from those of K&J for identifying a DNA descent cluster for a single descent family.

Some More Recent Y-DNA Developments

It is common in one-name studies to measure 37 Y-STRs for each man at the testing company Family Tree DNA (FTDNA): this is called a Y-DNA37 test. Moreover, FTDNA provide as many as 111 Y-STRs as a standard test called Y-DNA111. Beyond that, it has recently become more affordable to carry out comprehensive testing for novel Y-SNPs, using Next Generation Sequencing (NGS). FTDNA's so-called BigY test of Y-SNPs can determine further haplogroups by NGS and this test now costs less than twice that of a Y-DNA111 test of Y-STRs.

To illustrate some recent advances that have resulted from NGS determinations, Figure 3 shows a Network diagram[11] based on Y-DNA37 data for men that have been found to be in the rare haplogroup R-L617.[12] This haplogroup is a sub-clade of a parent clade R-DF27, which is part of R-P312, which is part of R-M269, which is part of R1b1, which then accounts for a substantial fraction of all men in Western Europe. By determining more Y-SNPs for a man, it is possible to identify more tightly his haplogroup (*i.e.* a more recent sub-clade) in the so-called human haplotree of male-line descent. Each descending sub-clade has arisen at a later time and progressively contains less of the world population.[13]

Though it is not a hard and fast rule, the more widely separated are the circles in Figure 3, the more likely their descent lines have diverged at a time near the earliest era of their haplogroup: around four millennia ago in this case of R-L617. It can be seen that the Plant cluster (individuals prefixed P) is well isolated from the other known individuals; this remains true when others not in this haplogroup are also considered[14]. On the other hand, those in the circles prefixed T have various surnames (Tyndal, Teague, Marsh, Spink and Westmoreland) and they nearly match with the Marsh descent cluster (individuals prefixed M). It is uncertain whether, for example, (a) these differently-surnamed individuals might have arisen through undisclosed Marsh paternities producing offspring in these other surnames, or (b) they all could have shared a male-line ancestor in the millennia that have passed between the formation of the R-L617 haplogroup and the development of English surnames.

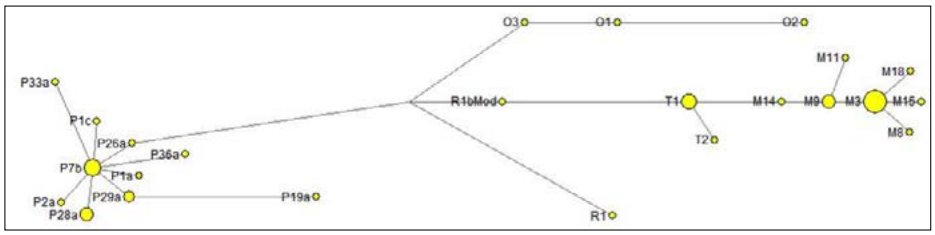

Figure 3: Living individuals belonging to the R1b-L617 haplogroup (R1bMod is the modal signature of a parent haplogroup R1b).

Large-Descent Families

For a large surname, there are typically many ambiguities of forename (e.g. many called William or John) in the documentary evidence. Our computer simulations also reveal the complications of how common surnames can contain very large as well as smaller descent families.

For a very large descent family, particular problems arise. Many bearers of the same surname are typically found near the large family's ancestral homeland. Even for those who are elsewhere, there are uncertainties about whether there are other descent families of the surname in that neighbourhood, or whether more than one branch of the very large family has migrated there. Each descent family can usually be distinguished relatively readily by straightforward Y-DNA testing, though this is complicated by NPEs. However, small Y-STR differences between different branches of a large descent family are generally much more difficult to identify and interpret.

For a single descent family to have grown abnormally large, some branching in the family can be expected to have occurred in the centuries soon after the family's medieval origin. We cannot rely on Y-STR markers to characterise the different branches. That is because these markers are typically fast changing, some more so than others. The same Y-STR mutation can have occurred down different branches of a descent family (parallel mutations); also, a fast-changing marker can mutate back to an earlier value (back mutation). There are accordingly ambiguities about the order in which the observed Y-STR mutations have occurred, making it difficult to ascribe one to a particular time in a particular branch. For example, a mutation might have occurred early in a particular branch followed in a sub-branch by a different marker mutation which however has also occurred as a parallel mutation early in a different branch.

More promisingly, not only are Y-SNP mutations much more stable, there are very many different ones that can arise. Unlike the Y-STR, it is very unlikely that there will be parallel or back mutations of a stable Y-SNP within the timescale of a surname. An NGS test can determine different Y-SNPs down different descent lines of a family, revealing perhaps half a dozen Y-SNPs that characterise each line.

We can hence much more safely identify an early Y-SNP mutation in a descent family, by seeking mutations that are shared by many of the family's men. This can identify a major branch of the family. Subsequent Y-SNP mutations, shared by fewer, can then characterise subsequent sub-branching within the branch. Upon reaching the times when adequate documentary evidence is available, we can combine Y-SNP, Y-STR and documentary evidence to identify still further sub-branching down to the multitudinous twigs of a large family's individual men.

We describe elsewhere[10] a case study where we identify four major branches of a large descent family, using just Y-STR data alone. This process can be expected to become simpler, more assured, and produce more detail with the growing availability of affordable and comprehensive Y-SNP data.[12] ■

References

[1] Plant, JS and RE (2013) JoONS, 11(7), 10-11.

[2] Plant, JS and RE (2014) JoONS, 11(9), 10-12.

[3] Plant, JS and RE (2014) JoONS, 11(12), 16-18.

[4] Plant, JS and RE (2014) http://www.one-name.org/GettingTheMost-Guild.pdf

[5] Plant, JS and RE (2014) http://www.one-name.org/ESDE-Guild-June2014.pdf

[6] Plant, JS and RE (2015) http://dx.doi.org/10.14487/sdna.001652

[7] King, TE and Jobling, MA (2008) http://www.ncbi.nlm.nih.gov/pubmed/19204044

[8] MEGA6 software, http://megasoftware.net/

[9] e.g. Kennett, D, *DNA and Social Networking* (2011, The History Press).

[10] Plant, JS and RE, http://www.plant-fhg.org.uk/dna.html#PlantMatching

[11] Flexus Engineering Network software, http://fluxus-engineering.com/index.htm

[12] Plant, JS and RE, and Marsh, AJ, http://www.plant-fhg.org.uk/dna.html#L617

[11] http://yfull.com/tree/R1b/