English Surnames: plural origins, DNA and emigration

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

For a wide range of surnames, we will outline a quantitative assessment of plural origins and emigration. This follows on from our earlier surname analyses (JoONS 11(7) pp. 10-11, and 11(9) pp. 14-16). First, however, we briefly digress on a minority view, that large surnames are single origin.

Family historian George Redmonds, peering through the mists obscuring medieval origins, offered guidance: given the almost unlimited choice of a name, re-using one was unlikely, implying surnames with only one origin. The geneticist Bryan Sykes used some early DNA data to concur that the Sykes surname is single-origin. However, if the Sykes name were ascribed only once, there are difficulties in explaining its large living population unless we allow, for example, that it could have been allocated to a few people, its once being at a single place.

Plant, also, might have been locational — its dominant family is found first near a newly planted vaccary. However, for a single beginning, much medieval migration would have been needed between distantly separated records of the name.

We could accept Richard Coates' view that a peasant did not choose his own surname — we can imagine a scribe who chose to use, in his records of the manorial court's proceedings, a succinct single surname for a small group of men. Such men could have shared local ancestry which would explain why DNA-matching populations are large in some common English surnames. However, this version of events relies not least on a hazy view of a scribe's predilections.

More traditionally, the belief is that prolific surnames such as Smith have plural origins, which arose in different places. We will assess this quantitatively below, to the extent that we consider that common and very common surnames, with UK populations over 10,000 or so, need to have had more than one 14th century

forefather.

Earlier, we examined some geographical distribution data and also made some predictions from our computer simulations. Either method suggested that it might be rare for a single-family population to exceed around 10,000 people in the UK. By a single family, we here mean a separate-origin descent from a single 14th century progenitor, experiencing observed overall growth conditions that no doubt reflect plagues such as the Black Death, conflicts, and regional economic differences.

Using regional growth factors that neglect the possible effects of migration, we simulated male-line family growth. Besides regional effects, our computer predictions reflect, in particular, random fortuity at each generation in the bearing of fertile sons who survive to father their own children. A particular prediction from the simulations is that a large observed DNA cluster does not necessarily mean that a surname had just one male progenitor: the simulations predict that most descents from other origins would have died out or grown relatively little, with sometimes a large descent family standing out from the rest.

Turning to some observed data for real

surnames, their populations are represented by the blue diamonds in Figure 1. These populations range from 148 (Feakins) to 775,645 (Smith). This Figure also indicates the size of the largest biological cluster, which can be taken to represent a descent family that contributes towards the surname's whole living population. Most of the underlying DNA data, for the red squares, comes from a study by the geneticists Dr Turi King and Prof Mark Jobling (K&J), though they themselves did not use their data in this way.

Most of the surnames in Figure 1 have populations below our predicted limit, which we estimated as 10,000 for a very large single family in the UK. It is hence quite to be expected that some of the smaller surnames could be single-origin and contain just one small or moderately-sized descent family. On the other hand, the whole population of the very common surname King, for example, exceeds this single-family limit, implying that it should have many other separate-origin descent families besides its largest.

Mostly, this is confirmed by the evidence in Figure 1. Many of the red squares for the smallest surnames leave little room for other small families besides the larg-

Figure 1: Largest cluster populations in some UK surname populations.
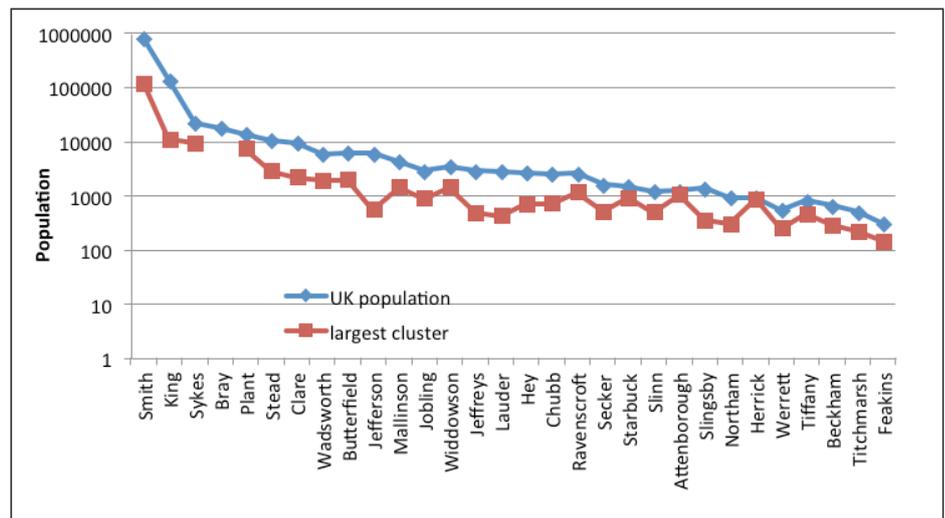
est, especially after allowing for the effects of non paternity events (NPEs). This is less true for some larger surnames, such as Jefferson. For King, its red square suggests that it has a largest biological descent family of around ten thousand, leaving room for many other families. However, the red square for Smith suggests a single family size of around a hundred thousand, far exceeding our predicted limit, though room still remains for many other families. We will proceed to offer an explanation for the Smiths' high red square and the absence of an observed matching cluster for Bray.

It is important to bear in mind some practicalities of the DNA determinations.

The red square for the largest descent family, for King or Smith, is derived from a small DNA cluster: a cluster of two in a sample of 24 in the case of King; and, a cluster of nine for Smith. Small clusters are sensitive to statistical bias and sampling error. For Smith, its observed 15^{1}/$_{2}$ per cent DNA cluster might arise from the limited DNA resolution used by K&J, which may not have distinguished adequately between separate medieval families for this surname. It is hence just dubious speculation that a particular family of Smiths might have had a very early genesis followed by an anomalously high male-line growth. That would be at odds with the predictions of our computer model simulations, at least with its simplest assumptions.

Another practicality arises in connection with the gap in Figure 1 for the fourth surname, Bray. It indicates that K&J found no observed DNA cluster in their sample of 29 men. This observation implies a limit of 7 per cent on the fraction that would match hypothetically in a larger random sample of this surname's UK population. This is similar to the 9 per cent fraction found for a larger sample of the smaller surname Jefferson.

We should add that it is not generally reported whether hobbyist's DNA samples are biased to over-represent particular families. Our analyses assume a random sample of the surname's whole population in a region, in particular for men living in the UK. The K&J data were obtained from a truly random sample, but other surname data are almost never obtained this way. Instead, to use the statistical term, the data are

obtained "haphazardly". This does not cause a serious problem provided that data values are obtained independently of each other. One-name researchers often target particular families in a surname, or sometimes deliberately target at least two in each genealogical family, regardless of each family's size. This can produce a statistical bias that is significantly misleading for our purposes.

Practical problems promote an approach of validating our computer model, by using some suitable data. In particular, we consider moderately-common surnames — ones with living UK populations of around ten thousand people. For these, as we have noted, Sykes and Plant appear to be at one extreme, displaying a substantial DNA matching cluster, as against Bray or Jefferson at the other, with only a small largest cluster. This can be compared with Figure 2, which shows some simulation results for the example of 2,000 reproductively-active men, corresponding to a moderately-common surname of around 8,000 to 12,000 people.

Starting with Figure 2(a), this displays the predicted chances of there being different numbers of surviving separate-origin families in a moderately-common surname's living population. The tallest blue bars illustrate that we can expect around thirteen separate-origin descent families in a surname that experiences the general population growth rates of England; there is some spread of uncertainty, due to random fortuity, as represented by the spread of the bars. Rather fewer descent families (orange bars in Figure

2(a)), with a spread around seven, are expected for a surname of this size, if its families are in a geographical region such as Staffordshire where the general population data indicate average growth rates that have been relatively high.

Moving on to Figure 2(b), this shows that the largest UK descent family in a moderately-common surname is likely to have around 500 reproductively-active men, if it experiences only the average growth conditions in England. This corresponds to the tallest blue bar. However, for the high-growth region of the orange bars, a single-origin descent family might grow as large as 1,500 or more such men, though the shorter bars indicate that this has a lower chance of arising.

In other words, Figure 2(b) predicts that the size of the largest descent family in a moderately-common surname will generally be below three quarters of the surname's population. This corresponds to a DNA matching fraction of below 0.45, provided that we assume a non paternity event (NPE) rate of 2 per cent per generation. If the NPE rate were assumed to be lower, the predicted matching limit would be lowered less from 0.75.

The observed largest DNA descent clusters (red squares) in Figure 1, for moderately-common surnames, largely confirm this prediction of DNA matching values below 0.45. For Sykes, Bray, Plant, Stead, Clare, Wadsworth, Butterfield and Jefferson, the fractions are 0.44, <0.07, 0.50, 0.28, 0.24, 0.33, 0.33 and 0.09 respectively. For the high value for Plant, it should be added that
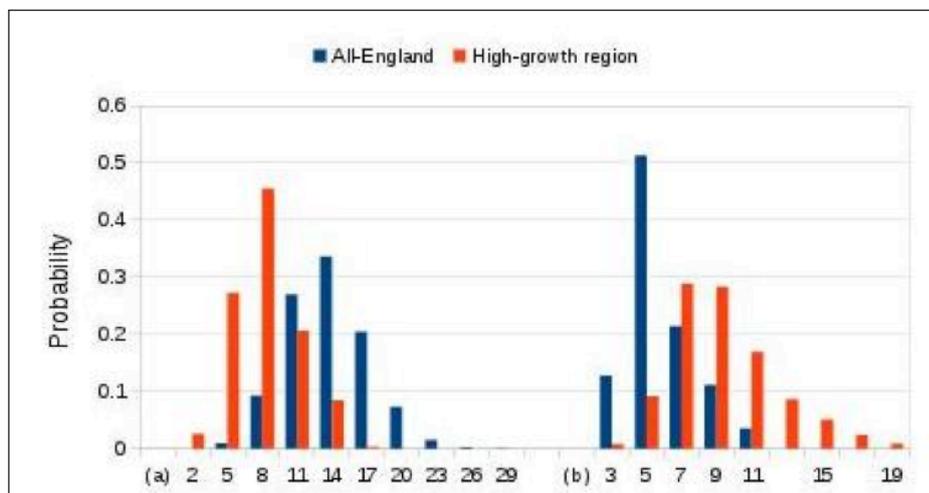

Figure 2: Predicted probabilities of (a) the number of separate-origin male-line families in a surname with 2,000 reproductively-active males; and, (b) largest family in hundreds of reproductively-active males.

there is a 0.12 standard error of statistical uncertainty arising from the limited sample size. It would seem that all of these surnames are plural origin though some evidently have a dominant family that accounts for a large fraction, perhaps around three quarters, of this surname's whole UK population.

We have also extended our computer model to include emigration. For this, we have derived, from published data, some historical rates of emigration assuming that it occurs randomly. However, it seems that there might have been a small non-random component to the way in which some surnames migrate. Non-random emigration could have arisen, for example, as follows. A surname's largest family in the UK might have grown abnormally and we can conjecture that this would have placed pressure on inherited land. Hence smaller, widely-spread English families might be expected to have had lower historical rates of migration than a large family experiencing land shortages.

DNA evidence suggests that there could have been non-random emigration of small and large families for the illustrative example of the plural-origin surname Plant. For this surname, the observed DNA matching fraction is 0.50 for a sample of sixteen men living in the UK, as against 0.71 for twenty-one men living overseas. Though this noticeable difference is not statistically significant at the 95% confidence level, it does not rule out non-random emigration. There could be several small descent families, lowering the DNA matching fraction in the UK. It could then be that the dominant descent family has emigrated disproportionately more, leading to a higher DNA matching fraction in the diaspora. Such non-random migration might show up in other ways.

For a wide range of surname sizes, Figure 3 presents some results from our computer simulations, which here assume purely random emigration. The simulated populations (grey circles) show that the overseas populations of the surnames increase roughly in line with their UK population. The superimposed coloured circles are for a few moderately-common real surnames (UK populations around 10,000). These surnames tend to lie towards the lower edge of the grey circles, indicating relatively low emigration rates. On the other hand, the triangles are for very common surnames (UK population around 100,000) which tend to reach more often towards the upper edge of the grey circles. This observed trend has been found for many real surnames. It might relate partly to our aforementioned comments on non-random emigration.
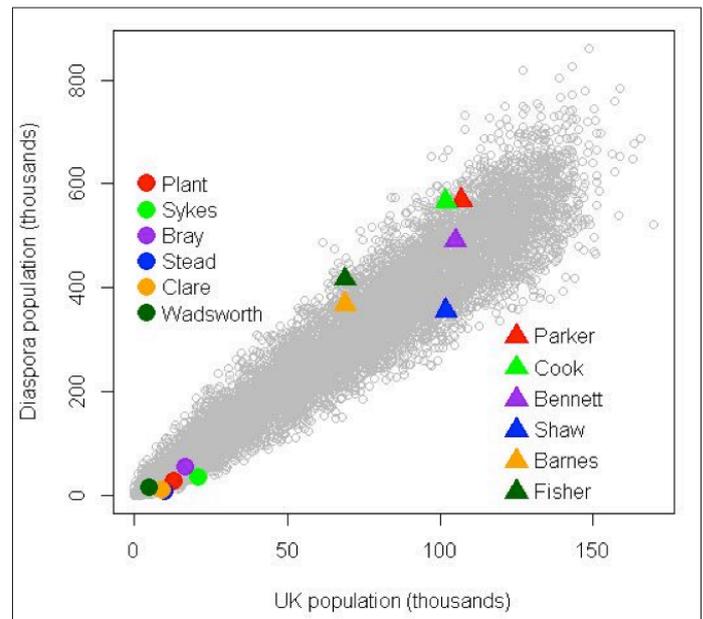
We give some further details at **http://www.one-name.org/ESDE-Guild-June2014.pdf** on the Guild website. ∎

Figure 3: Simulated relationship between UK and diaspora populations, with some data for real surnames superimposed.

DNA

to success. If you are pressed for time, do one step a day or one step a week. Once you have protected your registered surnames, you proceed at your own speed. Ideally, after you do your Getting Started email, your first priority is to recruit males for those trees with only one or a few surviving males, as well as test two distant males in your own tree.

6. You learn as you go and encounter different situations. This includes how to interpret results. In addition, the DNA Advisor is available for questions and consultation, and the Forum is available to post questions.

If any or all of your surnames are in an existing DNA Project, also write to the DNA Advisor. The existing project will be investigated, to determine the best course of action.

If an extraction of your surnames is the best course of action, you will be asked to write a short document or brief on your surnames. This is to provide information to be used in negotiations as to why your surnames aren't variants of the other surnames in the existing project. I have received many excellent briefs from Guild members, and it makes the extraction negotiations go fast, easy, and most importantly, successful.

If extraction is not an option, then a role is negotiated for you with the project. You decide what you want, and the DNA Advisor negotiates to achieve this outcome. Perhaps you only want to recruit UK persons, or you don't have time now, and want a role and access to the data in the project, and will provide suggestions or leads from your research to help the Project Administrator recruit.

Or do it on your own
You also always have the option of going forth on your own, whether it is setting up a project or negotiating a role in an existing project, or negotiating a surname extraction. ∎