# English Surnames: DNA, plural origins and emigration

by Dr John S Plant and Prof Richard E Plant

#### **Abstract**

Since the turn of the millennium, DNA testing has helped with establishing the sizes of biological descent families within surnames. Some surnames have a dominant family: their remaining populations can be ascribed either to NPEs (non paternity events) or to smaller separate-origin families. Previously, there was no direct evidence as to how large each male-line family could grow within a surname, considering that real-life plural-origin surnames could contain a few or many separate-origin families of various sizes. However, DNA testing is providing evidence to help indicate in particular the size of the largest male-line medieval descent family in a surname.

We here perform statistical analyses of some relevant DNA results. This allows us to compare the DNA results with some predicted findings from our computer simulations. We thereby surmise that moderately common surnames, with UK populations around ten thousand or so, are likely to be plural-origin, comprising a handful of separate-origin descent families. As limiting examples, Sykes and Plant are at the extreme of having a large DNA descent cluster, whereas Bray and Jefferson lie at the opposite extreme with only a small largest cluster.

We also proceed to consider very common and prolific surnames, with UK populations exceeding the order of a hundred thousand people. For these, the sizes of their descent clusters are fractionally very small and accordingly sensitive to sample bias. This places more reliance on the computer model predictions for such surnames. We model general trends for a range of surname sizes. For considering emigration overseas, we initially assume that it occurs randomly though we find some possible evidence that large descent families might migrate disproportionately more than smaller ones.

#### **Contents**

	Abstract	. 1
In	troduction	2
	Some terminology	
	Our extended computer model and comparisons with real surnames	
	Some UK DNA data	. 4

Some practical considerations relating to anomalous results	6
Some aspects of our previous computer simulations for England	7
Emigration	8
Observed emigration patterns	9
Computer simulations of whole surname emigration	13
DNA analyses	14
Suitable unbiased DNA samples	14
DNA matching into clusters	15
Largest families and calculated remnant fractions	16
Computer simulations for plural-origin surnames	18
Moderately-common UK surnames	18
Generalisation to other sized surnames	20
Largest predicted family sizes in moderately-common surnames	21
The small size of the largest cluster found for Bray	24
Different DNA matching fractions for Plant in the UK and in the diaspora	25
Computer simulations for very-common surnames	27
Appendix A: The Plant DNA results	28
Appendix B: Estimating the Properties of the Plant Population	33
B.1: Estimating the fraction of genetic descendants	33
B.2: Estimating the Modal Haplotype	34
Appendix C: The less certain data for Sykes	36
Appendix D: The Simulation Model	37

#### Introduction

We earlier presented computer simulations for the growth of populous single-origin families<sup>1</sup> and we here extend our computer model. We also broaden our deliberations, comparing our simulation predictions with observations for some real surnames whose component family populations are derived partly from previously published DNA data.

# Some terminology

We need first to clarify some terminology.

<sup>1</sup> John S Plant and Richard E Plant (2014) Getting the most from a surname study: semantics, DNA and computer modeling (third edition) <a href="http://www.one-name.org/GettingTheMost-Guild.pdf">http://www.one-name.org/GettingTheMost-Guild.pdf</a> and abridged extracts in the Journal of One Name Studies, 11(7) pp. 10-11 and 11(9) pp. 10-12.

So-called non paternity events (NPEs) give rise to a distinction between the biological and familial descendants of a medieval progenitor: the biological descendants involve a male-line genetic father passing on both his surname and Y-chromosome, whereas the familial descendants involve the transfer of the surname to children who are not necessarily true paternal descendants of a male bearer of the name. This is a key aspect of our interpretations of the DNA data which are obtained from DNA tests on the Y-chromosomes of living male bearers of the surname.

Only male-line biological descendants inherit the surname progenitor's Y-chromosome; his additional familial descendants inherit, typically through a female link, his surname but not his Y-chromosome. The additional component of familial descent in a medieval descent family of a surname arises because of, for example, the family adopting a differently fathered child, or the introgression of a differently fathered child into the family name through a concealed wifely infidelity, or an unmarried mother in the family passing on her own surname to her child of different paternity.

We shall here talk in particular about the "male-line descendants of a medieval man" or, in other words, his "male-line descent family". Considering either the biological or the fuller familial descent, we can consider that there could be several different contributing descent families that make up the total population of a plural-origin surname: we shall refer to each family as a "single-origin descent family" or a "separate-origin family". By this, we generally mean a surviving descent family from late medieval times whose single origin was when the surname first formed.

It is possible that several biological descent families will match at the level of resolution of the data, particularly when using DNA data that has only measured a limited number of Y-chromosome markers. As discussed further below, these several conflated biological descents can be indistinguishable from the assumption that a single large biological family makes up the DNA "descent cluster". By a "descent cluster", we mean a set of matching or almost matching Y-chromosome values. For each cluster, we can talk of its Y-DNA signatures which match or nearly match. A cluster is normally taken, when measured with adequate DNA resolution, to represent a single descent family but it might represent several, particularly if the cluster's Y-DNA signatures are common in the general population such that it might contain accidental matches to unrelated men.

Another ambiguity arises with Y-DNA, since it is difficult or impossible to distinguish between (a) small "separate-origin medieval descent families" and (b) descendants who have introgressed more recently with different paternity into a dominant single medieval family of the surname. Since the population of a moderately-common surname steadily increases down the centuries and since each individual in a descent family can be assumed to have the same chance of being subject to a non paternity event (NPE), most introgressions can be expected to have arisen through relatively-recent NPEs by when a dominant single descent family's population has grown larger than earlier. We can estimate the size of the full familial living descent population, descended from a single medieval male progenitor, by assuming a `reasonable value' for a constant rate per generation of the chance an NPE introgressing into the descent family's name. This enables us to apply a corresponding augmentation factor to the fractional size of the observed DNA descent cluster. We shall return later to tabulating the results of such calculations in Table 1.

# Our extended computer model and comparisons with real surnames

For the simulations, we have extended our computer model in three different ways.

First, we include the effects of NPEs (i.e. Non Paternity Events). We have considered them algebraically, separately from our computer simulations but we also include provision here for simulating them within our computer model.

Secondly, we extend our computer model to make use of published statistics on emigration, which allow us to simulate the emigration of bearers of English surnames to locations overseas where they may have enjoyed better growth conditions. This produces an overseas population which we will refer to as the diaspora. To determine parameters for the emigration model, we have derived the overseas growth rates from aggregated data for four populous UK surnames (Smith, Taylor, Brown and Wilson).

Thirdly, since we are concerned with some DNA results for men in the twenty-first century, we also extend our computer simulations up to 2001. Previously we simulated single-family growth from 1311 to 1881 only.

Fuller details about our computer model are given in Appendix D.

We will compare results from our extended computer simulations with some observed data for real surnames. When considering the analyses of some DNA data, we include, as a case study, some details of the analyses for our own surname. One reason for this is that we have particular knowledge of these data for Plant. For anyone wishing to apply our computer model specifically to their own surname data, we can supply our R-programming computer code. Where possible, we have also extended our deliberations to other surnames. For example, we have done this using some published DNA data for some surname bearers resident in England; such data are available for a range of different English surnames. We also consider population data for other surnames both in England and overseas. Such population data are readily available on the internet for nearly all English surnames although suitable DNA data for them is generally not as yet available.

#### Some UK DNA data

For Plant, we have in particular some DNA data for the UK and separately for overseas and we will compare, in this paper, these two separate datasets with results from our computer simulations. In general however, most of the currently available DNA data does not distinguish between test takers living in the UK or overseas. Nor does it mention the date when an overseas test taker's male line emigrated. Fuller analyses could become possible if more DNA project administrators clarified such details in their publicly-available datasets.

There is however a particularly useful published study for some bearers of English surnames though it is restricted to test takers living in the UK. Figure 1 accordingly illustrates some surname DNA data that is taken mostly from a study by King and Jobling (K&J).<sup>2</sup> This covers surnames of various UK sizes, ranging from 775,645 (Smith) to 148 (Feakins). The blue diamonds display the UK populations of these surnames on a log scale; these are in broad agreement, on this scale, with the red squares

<sup>&</sup>lt;sup>2</sup> T.E. King and M.A. Jobling (2009) `Founders, drift and infidelity: the relationship between Y-chromosome diversity and patrilineal surnames', Mol. Biol. Evol. 26(5) pp. 1093-1102. http://www.ncbi.nlm.nih.gov/pubmed/19204044

which are the surname populations as stated by K&J. To derive values for the blue diamonds, we have taken surname population frequencies for the UK from the *worldnames* website<sup>3</sup> and multiplied them by a value for the current total UK population, specifically by 63.7 million.

As already stated, the red squares in Figure 1 represent the surname populations as stated by K&J. However, since the surnames Sykes and Plant do not appear in the K&J study, the red squares shown for them are purely invented. Their blue diamonds are calculated as stated above and then these particular red squares are simply estimated to be in line with K&J's general underestimates of the blue diamonds. The other red squares are taken directly from K&J who state their numbers of surname bearers are for the year 1996 – their underestimations of the values corresponding to the blue diamonds for the UK can be ascribed only partly to the growth of the UK population since 1996.

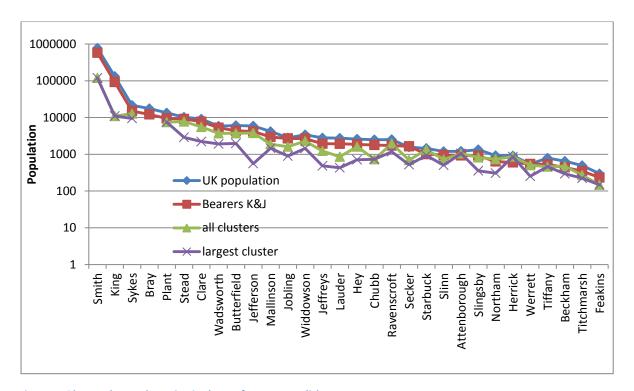


Figure 1: Observed DNA clustering in the UK for some English surnames.

Observed DNA data can provide us with a value for the fraction matching, or nearly matching, in a sample of UK volunteers. For this, care is needed to choose a random sample of DNA volunteers for each surname. In Figure 1, the observed matching fraction has been multiplied by the surname's UK population (blue diamonds) to give an estimated population, first, for the total of all of the surname's observed DNA matching clusters (green triangles) and, secondly, for the single largest matching cluster for the surname (purple crosses). This latter derived population is an indicator of the largest descent family in the surname. In fact, this matching cluster applies in each surname to the largest biological descent family, which can be expected to be rather smaller than the

.

<sup>&</sup>lt;sup>3</sup> http://worldnames.publicprofiler.org/Main.aspx

corresponding familial descent because the latter will include NPE introgressions. However, because of the logarithmic scale of the vertical axis, the expected extent of this difference is not expected to show up greatly in Figure 1.

In going from the surname Feakins to Stead in Figure 1, the total surname populations (blue diamonds) increase by a factor of 35 but the largest biological descent clusters in each surname (purple crosses) increase in size more gradually, by a factor of about 20. This suggests that the biological descent clusters begin to approach a limiting size and tend to level off. Particularly large biological descent families of around 10,000 are indicated in Figure 1 for King and Sykes (purple crosses of second and third surnames). Although the total UK population of King is around 6 times larger than that of Sykes, its largest biological family is only around 1.2 times larger in keeping with a supposition that the size of the descent clusters tend to level off.

However, the results in Figure 1 for Smith and Bray stand out as appearing to be anomalous.

#### Some practical considerations relating to anomalous results

The large size of the largest DNA cluster for Smith, shown in Figure 1, can be most readily explained as having arisen from a few separate biological descent families which accidentally DNA match nearly one another to form a large cluster of biological descent families with overlapping Y-DNA signatures. This phenomenon is sometimes referred to as `DNA convergence'. Specifically, it seems that the limited DNA resolution used by K&J, who measured only 17 Y-STR markers, did not adequately distinguish between some separate-origin families for Smith which apparently share near-matching Y-DNA signatures.

This large Smith cluster belongs to the common DNA haplogroup R1b and the Y-DNA signatures of this cluster could have been common in the surrounding general population. We hence do not need to consider that a single biological family has grown much larger than around 10,000 which is the apparent limit indicated by the other surnames. In other words, it would be no more than dubious speculation if we were to take the large cluster of near-matching Y-STR signatures at face value. It is only if we were to make such a presumption that we could interpret the observed DNA cluster as meaning that there was, for the Smith surname, a particular family of smiths with a very early genesis that has grown anomalously.

If that were the case that the descent cluster for Smith arose from a single descent family, then this presumed descent family would far exceed our predicted limiting size for most families that we expect from our computer simulations. That is to say, such a supposition of an anomalously large single descent family for Smith would be at odds with the most straightforward assumptions and predictions of our computer model.

At the other extreme, the lack of any apparent cluster for Bray can be ascribed partly to the limited sample size (29) of those men whom K&J recruited for DNA testing. Though the available results so far suggest that this surname is multi-origin with no dominant single-origin biological family, relatively small clusters of closely matching DNA signatures would no doubt emerge if the sample size of those tested were increased with more recruited random volunteers. These small clusters would be for relatively small descent families. In fact, there is some further DNA data for Bray, obtained using Family Tree DNA (FTDNA) testing but, as is generally the case with available FTDNA data, there are problems with knowing the precise quality of the data for our purposes.

The problems with the generally available FTDNA data are twofold. First, it is not clear which results are specifically for volunteers living in England. Secondly, it is not clear whether volunteers have submitted their samples independentlyor whether an observed DNA cluster is misleading, for our purposes. It is often the case that FTDNA data are for closely-related recruits who have been deliberately targeted as being quite closely related; this is often done for the different purpose of scrutinizing a specific male-line genealogical family tree. That would imply that the sample of volunteers is not representative of the surname as a whole but that it is instead biased to over-represent particular families regardless of their individual sizes, rather than randomly sampling the full population of the surname in a specified geographical region.

The lack of any observed matching clusters in K&J's random sample for Bray in the UK suggests that either it has undergone many NPEs or that this surname had many distinct medieval origins. In his Surname Dictionary, P.H. Reaney suggested various derivations for Bray: (1) from Bray in Berkshire, High Bray in Devon, or some unidentified place in Cornwall; (2) from the Cornish "bregh" meaning fine or brave; (3) from the female name Breya which might derive from the Cornish nickname; (4) from one of the many places in Scotland called Brae; (5) either a toponymic in Ireland or from a native of Bregia. It would hence appear that Bray could have had many separate origins throughout the UK. Its observed widespread geographical distribution in England reinforces this suspicion. We shall consider computer simulations for the plural-origin nature of a surname, such as Bray, later.

# Some aspects of our previous computer simulations for England

We have discussed the apparently large sizes of the dominant biological descent families for Sykes and Plant (third and fifth surnames in Figure 1) in previous publications.<sup>4</sup> The DNA data for Sykes is less clear cut than was initially presumed (Appendix C). With this reservation, these two surnames provide specific examples of ones that can be compared with our earlier computer simulations and they thereby evidently correspond to an extreme of large single-family growth.

In our computer simulations, there were very large effects from random fortuity, whereby sometimes, though not commonly, there can be relatively many reproductively-active sons in many generations of the family. Another important factor in determining the computed results is typical family growth in a particular region. There appear to have been different average growth conditions in different regions of England and some regions evidently provided relatively favourable conditions. This is indicated by the overall population changes down the centuries in a particular region which seem likely to relate at least partly to particular local conditions and not just to migration between regions.

Previously, we also quantified the ways in which the chances of large family growth could be enhanced by other factors. For example, we considered such features as, on the one hand, early polygyny, or, on another, an early start to a surname family's population.

Referring first to our model of "early polygyny" (literally meaning many women), there are such possibilities as that of a relatively high status peasant family successfully attracting more than one

<sup>&</sup>lt;sup>4</sup> http://www.one-name.org/GettingTheMost-Guild.pdf and abridged extracts in JoONS 11(7) and 11(9) as previously detailed.

successive fertile wife for each man, over a few early generations of the surname. This could have a significant effect on the size that a separate-origin family could attain. A population advantage gained in early generations would carry through as a firm basis for the growth of the whole descent family in subsequent generations.

Then, for our "early start" model, there is such a possibility as that of a much earlier man having had several male descendants who survived at a single location and who accordingly produced a local dominance of a particular male-line Y-DNA signature. Many of these male related descendants could then have been ascribed the same surname. This could apply to a topographical surname that arose from the name of a topographical feature at an isolated location, for example.

To elaborate on this "early start" possibility, we can note that separate computer simulations can be applied to pre-surname times and they show that a few separate-origin families at an isolated location will have tended to have experienced population drift over several generations towards having one dominant male-line biological family in the locality. In other words, most of several families at an isolated location would have tended to die out in the male line, leaving mostly one dominant descent family there in late medieval times.

This allows that a surname allocated to a few men at an isolated location in (say) the fourteenth century could apply mostly to men who had the Y-DNA signature of a single pre-existing "early start" male-line family. More generally, a topographical surname might have been applied to several unrelated male-line men, as well as related ones, at a designated place. Alternatively, a locative surname might have been given to several men who had migrated elsewhere from a named place.

Surname allocation to many unrelated men can be expected to lead to less male-line DNA matching and smaller and more observable fractional DNA descent clusters, genetic drift notwithstanding. Possible implications of genetic drift, however, include that most of the unrelated descent families could have died out.

Surname allocation to related men might lead to a relatively large DNA cluster provided that they were a family that fortuitously went on to have repeatedly many fertile sons.

# **Emigration**

Although there are only a limited number of suitable DNA studies to provide data for helping with statistical studies of single-family emigration, we can hope that more suitable DNA data will be forthcoming as people better understand the potential of such data and the requirements for assembling a useful and unbiased dataset. We will here make a start however, using non-DNA data, in delving deeper into the statistics of separate-origin family emigrations in the context of populous surnames.

We start with considering observed population data for whole surname emigrations, leaving aside for the moment the detailed situation for the separate families within them. We shall return to a consideration of the latter towards the end of this article. The population data for whole surnames in various countries are here taken from the website *worldnames.publicprofiler.org*.

# **Observed emigration patterns**

As examples of whole surname emigrations, we can consider a few moderately-common ones such as the third to eighth surnames in Figure 1. For these, there is well documented DNA data for the UK though, for the moment, we will mention that data only in passing.

Figure 2 illustrates the emigration patterns for six moderately-common surnames from Figure 1. These initially chosen surnames are ones for which there are published DNA data<sup>5</sup> for male bearers of the surname in the UK. These surname bearers have either volunteered independently or they have been invited at random to participate. Their reported fractions that DNA match in the UK, to form a dominant biological descent family, are respectively 44%, 0%, 50%, 28%, 24% and 33%.

The 0% DNA matching for Bray has already been mentioned. Taking account of the limited size of the tested sample, the matching is in fact measured to be less than 7%. This can be taken as one of a few indications that this surname is multi-origin. In other words, this surname appears to comprise many small UK families from distinct medieval surname origins, though some of the mismatching no doubt arises from more recent NPE introgressions into the surname.

Returning to whole surname considerations, the red bar for Bray in Figure 2 indicates that a sizable fraction of its living population is now in the USA.

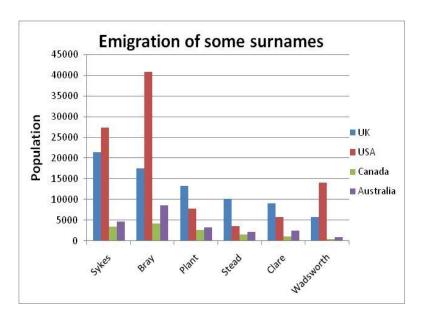


Figure 2: Some fairly large English surnames and the main destinations of their emigration overseas.

-

<sup>&</sup>lt;sup>5</sup> Bryan Sykes and Catherine Irven (2000) `Surnames and the Y chromosome', Am J Hum Genet 66(4), pp. 1417–1419. Turi King and Mark Jobling (2009) `Founders, drift and infidelity: the relationship between Y-chromosome diversity and patrilineal surnames', Mol. Biol. Evol. 26(5) pp. 1093-1102. Also, http://www.plantfng.org.uk/dna.html.

For the purposes of whole surname emigration however, we do not need to restrict our considerations to surnames with suitable available DNA data. Figure 3 plots, for many more English surnames, the relationship between their overseas and UK populations. The population data are taken from the aforementioned website. Broadly the population of their overseas diaspora increases in line with their UK population, although there is less of a pattern for surnames with smaller populations.

In Figure 3, the general trend of the plotted data does not pass quite through the graph's origin. There is an apparent excess in the fraction remaining in England for the smaller surnames in comparison to the larger ones. This will be evidenced further in Figure 4. The offset from the origin of Figure 3 might be ascribed to a number of different factors, such as:

- small surname populations might be under-represented in the overseas diasporas since, for example, we have included surname spellings in the UK and these might have changed to different spellings overseas;
- we have not included in the chart surnames with an overseas population for which the population has died out in the UK;
- eldest sons might have tended to stay at home disproportionately, on inherited land, and
  this is relevant since smaller families can be expected to have had a higher proportion of
  eldest or only sons; these families are hence more likely to have remained on inherited land
  and presumably to have been less likely to migrate; or,
- we might have selected, for the Lancashire and Staffordshire surnames, relatively immobile
  families in as much as we have chosen them as relatively tightly clustered surnames in those
  counties in 1881; this is different from the situation for the highest population surnames,
  which are spread more widely throughout England.

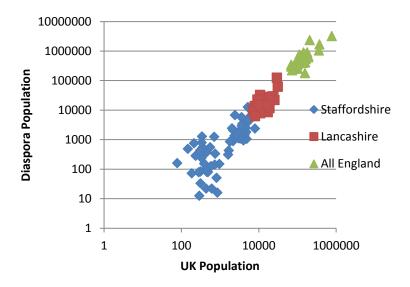
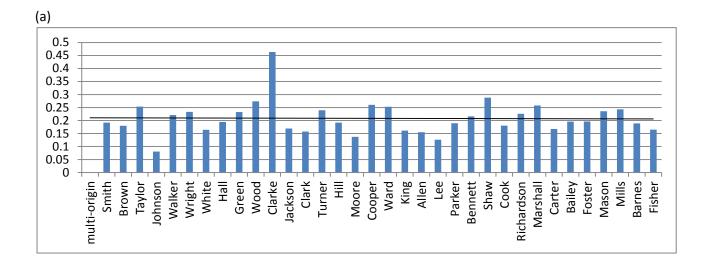
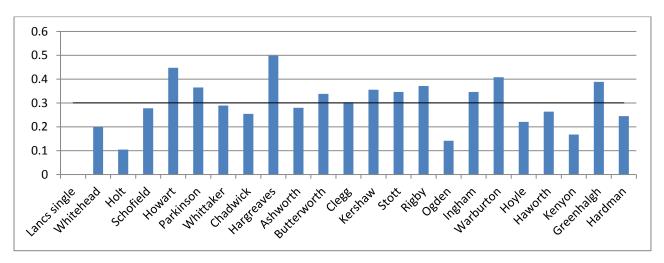


Figure 3: Extent of overseas emigration for various real surnames clustered in Staffordshire or Lancashire or spread throughout England



(b)



(c)

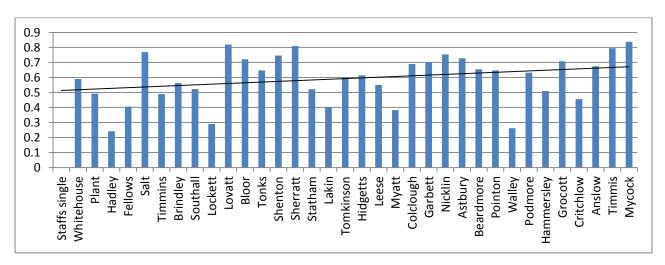


Figure 4: Fraction in UK for surname s that are: (a) widely-spread; (b) mainly in Lancashire; (c) mainly in Staffordshire.

The trend-lines in Figure 4 indicate more clearly the nature of the offset in Figure 3. Figure 4 shows typical values for the fraction of a surname remaining in the UK. This fraction, as indicated by the trend-lines, increases as the population of the surname decreases. In other words, as for Figure 3, it would seem that smaller surnames tend to remain more in the UK. This is apparent even when the surnames that are being compared are all in the same county.

The typical UK population fractions of the Figure 4 trend-lines are respectively: (a) around 0.2 for prolific and very common widely-spread UK surnames; (b) around 0.3 for moderately-common surnames clustered in the county of Lancashire; and, (c) around 0.5 rising to 0.7 for progressively smaller surnames in Staffordshire.

#### Computer simulations of whole surname emigration

Figure 5 presents, on two different scales, some predictions for emigration obtained from our extended computer simulations. The simulated populations (grey circles) are superimposed with a few results for real surnames. The simulated predictions are generated from 100,000 Monte Carlo simulations using a model with separate UK and diaspora growth conditions. The UK conditions have been taken to match the average for England in Figure 5(a); or, for Figure 5(b) the higher growth conditions found in Staffordshire. The diaspora growth conditions have been taken throughout to match a weighted average for English surnames in the USA, Canada, and Australia.

The simulations in Figure 5(a) are carried out assuming that the surname population is made up of 150 independent separate-origin descent families. This many is needed to cover the populations of the larger real surnames (triangles). More vigorous descent family growth than this, due to either local conditions or other favourable factors, would reduce this number of families required in the simulations.

For unusually high growth families, unusually favourable growth factors might apply, rather than family growth being just purely fortuitous within typical conditions for England. Unusually high growth could reduce the number of descent families required to cover the populations of very common surnames (coloured triangles in Figure 5(a)).

The coloured circles in Figure 5(a) correspond to the real surnames that were included in Figure 2. Although these surnames are much smaller than the surnames corresponding to the triangles, they are still for moderately-common surnames.

The grey circles from the simulations for Figure 5(b) assume a population comprising seven separate-origin descent families. The superimposed coloured circles here are for real surnames that might have a single or a few origins. In Figure 5(b), the superimposed triangles are for surnames that are more widely dispersed in the geographical distribution maps of Steve Archer's 1881 Surname Atlas. The coloured circles apply to less widely dispersed surnames. Assuming seven distinct origins (grey circles in Figure 5(b)) is fully enough to cover the observed populations of these moderately-common surnames.

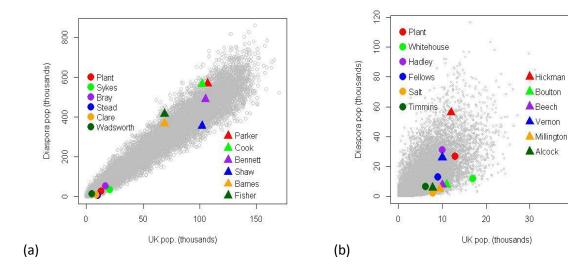


Figure 5: Simulated relationship between UK and diaspora populations, with some data for real surnames superimposed.

# **DNA** analyses

Turning to our analyses of some DNA data, we will here just briefly outline some key aspects. Fuller details are given in the Appendices.

40

#### **Suitable unbiased DNA samples**

Statistically, there is a standard error in taking the fraction of Y-STR matches in an unbiased random sample of a surname and assuming that the same fraction will apply to the surname's total living male population. The mathematical formula for this is given in Appendix B.1. The standard error should not be confused with additional systematic errors arising from sample bias. There is currently particular confusion amongst hobbyists concerning an associated factor of penetration.

Amongst amateur geneticists, there is some misinformation about what constitutes an adequately sized DNA sample. It has been misleadingly stated, for example, that a minimum value for a "penetration factor" is needed, with that penetration factor defined such that this would imply a need for a massive DNA sample in the case of a prolific or common surname. This needs to be debunked. Though the required sample size depends on how the sample is to be used as well as the variability of the data, it has been stated by one professional statistician, for example, that a size above 20 is large and one above 30 is extravagant. Rather than a misleading emphasis on a "penetration factor", it is of greater concern, for our purposes, that there is typically a major error amongst non-professionals in advocating a DNA dataset that shows disregard for the separate issue of sample bias.

-

<sup>&</sup>lt;sup>6</sup> Michael J Cawley (2005) *Statistics: An introduction using R*, ISBN-10: **0470022981** | ISBN-13: **978-0470022986.** Some similar issues also are considered by one of us in Richard E Plant (2012) *Spatial data Analysis in Ecology and Agriculture*, ISBN-10: **1439819130** | ISBN-13: **978-1439819135.** 

The term "penetration factor" has been used by hobby geneticists for the size of the number DNA tested men relative to the total male population of the surname. That fraction has some relevance for rare surnames in as much as, for small sized surnames, it is a dominant term in the calculation of the statistical standard error. For moderately-sized surnames however, the only significant term for the standard error is purely the sample size, irrespective of its size relative to the whole population of the surname.

Confusion should be avoided between sample size and a more professional usage of the term *penetration*. In the early days of the internet, for example, it was doubted whether those who volunteered to complete on-line surveys were representative of the general population as a whole; it was hence considered desirable to obtain data that *penetrated* more of the general population. For our purposes, it is similarly important that the DNA sample is representative of the whole surname population that is being considered and that it does not just represent an untypical subset. That is a matter of statistical bias: the sample should not be biased to over-represent just an untypical part of the whole surname population. This is quite separate from the statistical standard error.

It is important to draw a distinction between rare and more moderately-sized surnames. In general, statistical idiosyncrasies (cf. Figure 3) can be expected to arise for rare surnames. These have relatively small populations. Very diverse behaviours can arise for them because the full surname size is too small to show a distinct behaviour that can be associated with typical behaviour within the general population. For relatively rare surnames, the nature of a DNA investigation is generally quite different from that which we consider mostly here, for more populous surnames. Observed DNA results can be obtained far more readily for a large fraction of the male population of a rare surname and considered along with the genealogical family (or families) of all the individuals in the surname. In contrast, general statistical trends can be expected to be more pronounced and less idiosyncratic for more common surnames, as is illustrated in Figures 3 and 5 for example. For more moderately sized surnames, general population trends are more relevant and it is accordingly informative if an unbiased random sample can be obtained for the surname; this can be achieved by considering only independent volunteers for a DNA test from the whole population of the surname

As already mentioned, it is important to avoid biasing the sample by the over-selection of subjects from one particular subset of the population. That is unless, for example, one is deliberately considering the surname's characteristics in one specified region. In the case of worldwide DNA sampling, it is possible that a disproportionately large number of individuals might, for example, be from one country. In hobbyist DNA studies, it is quite typical for test-takers living in some part of the USA for example to be haphazardly over-represented. For the Plant DNA dataset, we have sought to avoid statistical bias. For example, chi-squared analysis shows that there is no significant bias in the sampling of Plants from different countries of the world (Table 2 in Appendix A).

#### **DNA** matching into clusters

In deciding whether individuals' results match into a primary DNA cluster, within a random unbiased sample, it is appropriate to include regard for a control sample. Taking our analyses of our Plant DNA results as an example, this available sample had a main DNA cluster for which its Y-STR results did not match the result for anyone else in the control sample, for which we used the large FTDNA database of test takers with many different surnames. This FTDNA control sample of results for all

test takers indicates that it is unlikely that the main Plant cluster has arisen from *accidental* matches within the surname. Even so, borderline Y-STR matches to the primary DNA cluster of Plant were confirmed by additional Y-SNP testing of a shared haplogroup<sup>7</sup> to give a combined probability of DNA match determined by a Bayesian calculation. K&J used a different approach for deciding whether nearly-matching randomly-recruited individuals belonged to the same DNA cluster. Further details of our approach are given in Appendix A.

When the control sample indicates that there is a significant chance of fortuitous matching, the problem is generally alleviated if more Y-STR markers are measured and rarer haplogroups are identified by testing additional Y-SNPs. Amateur geneticists, who proffer advice about what is necessary, often omit to add that the requirement depends very much on the control sample and the precise Y-DNA haplogroup, haplotype and locality of each genetic family that is being considered. It is simpler to proffer general guidelines that apply to the majority of cases.

The observed fraction of the Plant sample that Y-DNA matches into a dominant cluster is 0.5, with a standard error of 0.125, for volunteers called Plant currently living in the UK. It is higher at 0.76, with a standard error of 0.09, for the Plant diaspora. This suggests a possible prevalence of more DNA *mis*matching in the UK than in the diaspora. We will discuss this observed feature in a later section of this article.

In general, as already mentioned, observed Y-DNA results do not distinguish between mismatches due to (a) separate medieval origins to a surname and (b) those due to NPEs. We might hence seek further clues from documentary sources. In the case of Plant, we can note that most of the DNA matches in the UK are geographically concentrated, as expected for a random sample, around the main concentration of this surname as a whole, which is mostly around the county of Staffordshire. This is illustrated in Figure 17 in Appendix A for example. On the basis of the geographical evidence, we might conjecture that those Plants in England who are more distant from their Staffordshire homeland might be more likely to be descended from separate medieval origins: there is just some limited evidence to support this in both the medieval documentary evidence and the available DNA sample of UK results, when they are combined with some genealogical information. However, dispite this likelihood of at least one other separate-origin family, it cannot be ruled out that the living Plant population is exclusively, though more likely predominantly, a single-origin descent family.

#### Largest families and calculated remnant fractions

For the observed UK matching fractions, we can widen our considerations from just the case of Plant to include those surnames included in Figure 2. Surnames of that size can be termed `moderately common' and they are likely to be plural origin with more than one surviving descent family. As already noted, with the exception of Bray (<0.07), these surnames in England have DNA matching fractions in the range 0.25 to 0.5. We can take these values to indicate the fraction of the total UK population of each surname that forms a dominant male-line biological descent family.

\_

<sup>&</sup>lt;sup>7</sup> Haplogroup structure is based on a tree of Y-SNPs and this is currently a fast developing area. At the time of writing, the Plant haplogroup is adequately defined by R-DF27+, Z196-, Z225-. This represents the test results for just three SNPs. Many more SNPs have been tested but just these three define the position of the Plant haplogroup in the current experimental haplotree.

To this biological descent, we can add those who are associated with the same dominant descent family but who have obtained the surname through adoption or through a female link (i.e. through NPE introgressions). This will contribute to the total "familial descent" of the largest surviving family from a medieval progenitor. We can calculate algebraically<sup>8</sup> the expected size of this familial descent from the size of its DNA estimated biological descent – to do so, we need to assume a value for the NPE rate as well as for the number of generations of descent.

In Table 1, we assume 25 generations of descent from the most prodigious medieval progenitor. We also consider NPE rates of 1% and 2% per generation. If the dominant family is large (as suggested for some surnames by the purple crosses in Figure 1), it is reasonable to assume that it has grown through relatively many generations since medieval times, in order to reach its current large size – in our computer model, we assume descent through 25 generations, rather than a more recent development of the name and an ensuing descent through just a few generations.

Matching Fraction	.25	.3	.35	.4	.45	.5	.55	
Familial descent (1% NPE)	.32	.38	.45	.51	.58	.64	.71	
Remnant (1%)	.68	.62	.55	.49	.42	.36	.29	
Familial descent (2% NPE)	.42	.5	.58	.67	.75	.83	.92	
Remnant (2%)	.58	.5	.42	.33	.25	.17	.08	

Table 1: Remnant fractions of a plural-origin surname ascribable to separate-origin medieval descent families.

As an illustration of the implications of Table 1, we can assume that the observed DNA matching fraction for a surname is one of those values given on the top row. For example, a surname with an observed DNA matching fraction of 0.4 can be seen, in that column of the table, to have a calculated dominant family of familial descent amounting to a fraction 0.51 of the plural-origin surname's whole population, if we assume an NPE rate of 1% per generation. The fraction can be seen to become 0.67, if we assume a 2% NPE rate, leaving a third of the surname remaining —hence, the "Remnant (2%)" row in Table 1 shows the calculated value 0.33 for this column. This remnant is available to be divided amongst a number of smaller families that have descended from separate medieval origins.

Figure 6 graphs the full range the calculated remnant fractions taken from Table 1.

<sup>&</sup>lt;sup>8</sup> cf. Appendix A of http://cogprints.org/6595/.

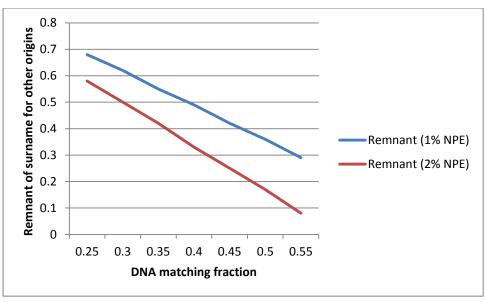


Figure 6: Calculated fraction of a plural-origin surname that is available for other small descent families.

# **Computer simulations for plural-origin surnames**

Some details of our computer model are given in Appendix D. We here outline some relevant computer simulation results.

#### **Moderately-common UK surnames**

We can consider the number of randomly selected families for our computer simulations that are needed to fill the population of a moderately-common surname. As a development of this, we consider the revised number if we have DNA evidence indicating the size of the largest family so that we can estimate how many additional families are needed to fill the remnant population. We begin here by ignoring any DNA evidence.

Based on our computer simulations, Figure 7 shows the predicted chances of there being different numbers of separate-origin families in the living population of a plural-origin surname. It applies to a surname with a UK population similar to that of the surnames in Figure 5(b); to wit, a moderately-common one with a UK population of two thousand reproductively-active males, which corresponds to a total UK surname population of around eight to twelve thousand people.

The bars in Figure 7 show the computed probability distributions for there being different numbers of families. These families have fortuitously varying sizes within the surname. They are randomly selected from the surviving families arising in Monte Carlo computer simulations.

The two different distributions in Figure 7 correspond to two different sets of historical population growth rates for the families. The blue bars assume that the growth rates match the observed historical rates averaged for all England; whereas, the red bars correspond to the far more favourable average growth rates found for the population of Staffordshire.

As one would expect, the more favourable Staffordshire growth rates lead to the need for fewer families in order to fill the whole target population of 2,000 reproductively-active males. The modal

number of families predicted with the Staffordshire rates is seven (red bars) whereas it is thirteen for the lower all-England rates. There are quite wide spreads around these modal values whereby the simulations predict other possible numbers of families in the surname; this spread arises from random fortuity in the bearing of fertile sons, as simulated in our computer model.

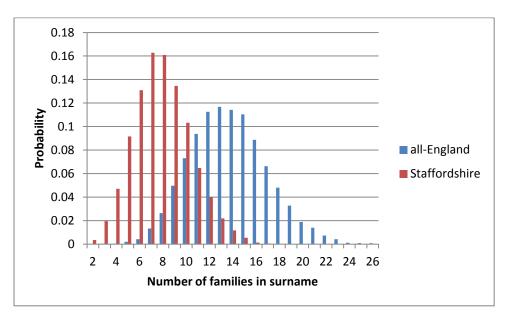


Figure 7: Predicted number of single male-line families in a surname with 2,000 reproductively-active males assuming the observed overall population growth rates of England or Staffordshire.

As a different estimation, Figure 8 feeds into the simulations an observed DNA matching fraction of 0.4 and two different assumptions for the NPE rate. To supplement the dominant family implied by the 0.4 DNA matching fraction, a number of extra families is drawn randomly from the simulations to fill the remnant population of the surname. The size of this remnant population is as given in Table 1 and illustrated in Figure 6.

In Figure 8, the extra smaller families are assumed to have experienced the all-England growth rates; and, the total population of the surname is again taken to be 2,000 reproductively-active males. It can be expected that fewer families will be needed for a higher NPE rate, since this leaves a smaller remnant population (cf. Table 1) for the extra families to fill. It can be seen that the modal number of families for the 1% NPE rate is six (for the blue bars) and it is fewer at four for the 2% rate (red bars). These numbers have had one family added to account for the largest family itself.

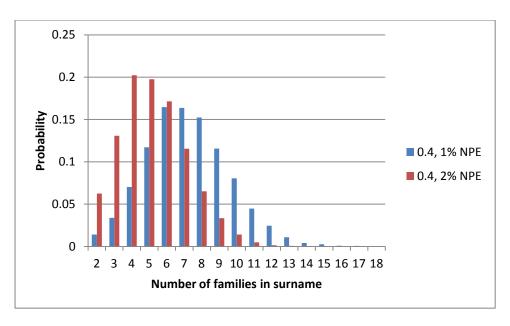


Figure 8: Predicted chances of different numbers of families for an observed DNA matching fraction of 0.4 and different assumed NPE rates.

#### **Generalisation to other sized surnames**

We can approach the simulations a different way. We can assume a specific number of descent families in a plural-origin surname and compute a range for the whole surnames populations. We can red Figure 9 in two different ways. We can select a number of descent families from the horizontal axis and look up vertically to see the range of predicted populations. Alternatively, we can select a surname population from the vertical axis and look across horizontally to see ar range for the predicted numbers of separate-origin descent families. For example, for a moderately-common surname we can select a surname population of around ten thousand. Though the computer simulations deal in the number of reproductively active males, we have here assumed a multiplier of six to scale up to the whole living UK population predicted for the surname.

Figure 9 accordingly illustrates a generalization of the findings of Figure 7 to other surname sizes. There is an 80% chance that the surname's population will fall between the computed 10<sup>th</sup> percentile (10pc) and 90<sup>th</sup> percentile (90pc) in the probability distribution of the surname's predicted size. The graph also displays the lowest predicted surname size with the all-England growth conditions (dark blue squares) and the highest predicted surname size in the high-growth Staffordshire conditions (light blue triangles). Usually, the surnames populations can be expected to be between the "England 10pc" and "Staffs 90pc" markers. However, for surnames in particular with just a few descent families, it is predicted that there can be a much wider range of predicted populations even than this, as indicated by the "England low" and "Staffs high" markers.

If there is just a single descent family (extreme right of Figure 9), the "England 10pc" marker predicts that, in all-England growth conditions, the medieval descent family will usually have a living population of at least 100. This is around a quarter of an average surviving family size. However, there are a number of reasons why a rare surname might be smaller than this: the surname might have arisen as a statistically rare event and survived with a smaller population without dying out; or, it might have developed under less favourable growth conditions than those typical of all-England;

or, it might not have formed as early as 1311 such that it has had less time to grow. As examples of the last possibility, the name could be a relatively recent invention or it could be derived from a recently arrived immigrant name.

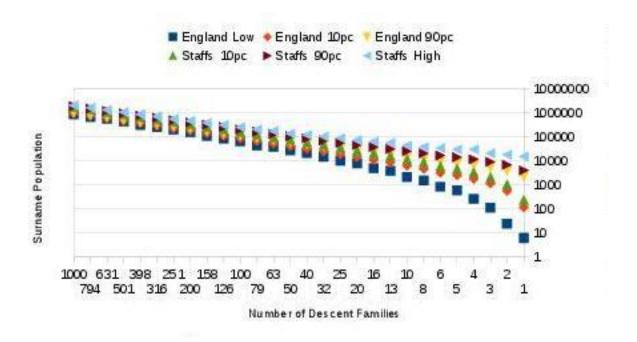


Figure 9: Predicted surname populations for different numbers of descent families in the surname.

#### Largest predicted family sizes in moderately-common surnames

Figure 9 places some limits on the predicted number of descent families in a surname. We can also compute predicted sizes for the largest family, as shown for moderately-common surnames in Figure 10(b).

Figure 10 applies moderately-common surnames in a region of favourable single-family growth conditions. Both Figures 10(a) and 10(b) apply to Staffordshire which, according to observed population data, is evidently a high-growth region in England. The blue bars in Figure 10(a) correspond to the number of separate-origin descent families that are needed, when randomly selected from the simulated descent families, to fill a target population of 2,000 reproductively active males. As before, this corresponds to a UK population surname size of around eight to twelve thousand living people. The orange bars in Figure 10(b) display the corresponding computed probability distribution for the predicted sizes of the largest descent family within the surname. This computation assumes that there are specifically seven descent families that are randomly selected from the simulations. It can be seen that the most probable largest family size is around 800 reproductively-active males which is 0.4 of the total surname size.

For an example of a real surname, we can consider the available DNA results for Plant. For this surname, we have an observed matching fraction of 0.5 for the UK though this applies only to the biological descendants of a single male progenitor to which an unknown number of descendants that involve an NPE should be added. If we assume for example a 2% NPE rate, we can estimate

(Table 1) that around four fifths or more of the Plant surname might descend from a single origin. This would correspond to a family size of around sixteen hundred or so in Figure 10(b). Though this estimated family size of sixteen hundred reproductively-active males is predicted as feasible by the short orange bars in Figure 10(b), it is predicted to be less likely than a familial descent of around eight hundred (tallest orange bars in Figure 10(b)). We can hence regard Plant as being towards the large single-cluster extreme for moderately-common English surnames.

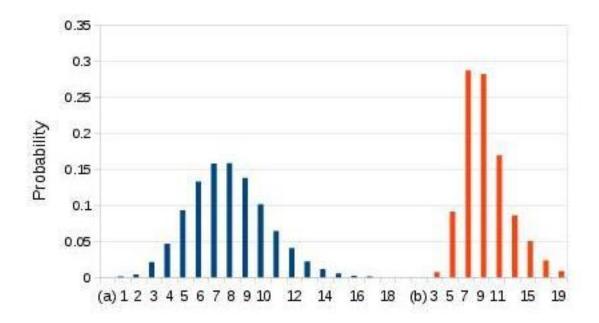


Figure 10: Predicted probabilities, for Staffordshire growth rates, of (a) number of separate-origin male-line families in a surname with 2,000 reproductively-active males; and, (b) largest family, in hundreds of reproductively-active males, assuming seven descent families.

The significance of the colours is different in Figure 11. The orange bars apply to the high-growth region of Staffordshire. The blue bars apply to the average growth rates for all-England. Comparing the orange with the blue bars indicates that, in the higher growth rate geographical region, there is more probability of the surname having fewer descent families (orange bars in Figure 11(a)); and, also, its largest such predicted family has more chance of being larger (orange bars in Figure 11(b)). Considering just the all-England growth rates of the blue bars, these indicate that there is a reasonable chance of 20 families arising in the simulations to make up the total surname population; however, there is a higher chance of only 13 families.

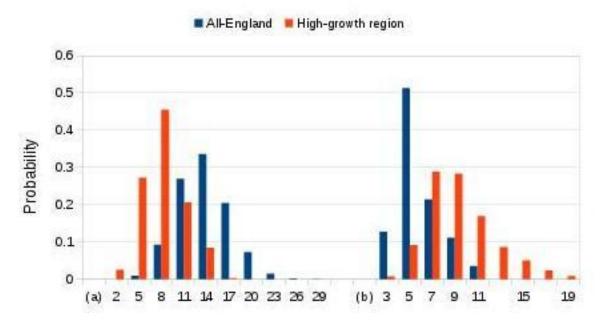


Figure 11: The effect of growth rate on (a) the number of predicted families amongst 2,000 reproductively-active males; and, (b) the predicted sizes of the largest family.

Accordingly, moving on to Figure 12, the blue bars correspond to 20 descent families making up the whole surname, assuming the all-England growth parameters, and the red bars correspond to the higher chance of there being only 13 descent families. It can be seen that ,with fewer families (red bars), there is less chance that the largest descent family will have only 2 or 3 hundred reproductively-active males, as illustrated by the shorter red than the blue bars towards the left-hand tails in Figure 12. In other words, when there are fewer descent families, this decreases the chances that the largest descent family will be towards the lower size extreme.

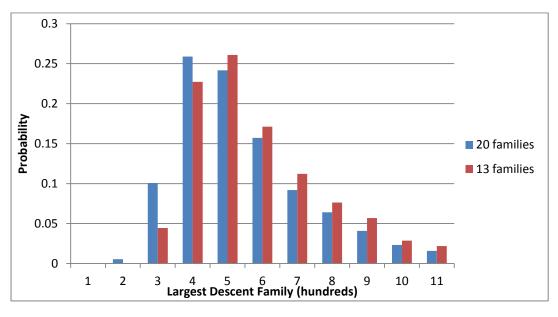


Figure 12: Largest predicted family sizes for 2,000 reproductively men made up of 13 or 20 families, assuming the all-England growth rates.

#### The small size of the largest cluster found for Bray

We have previously considered the relatively high matching fractions of the two moderately-common surnames Sykes and Plant. Sykes has a UK population of 21,562 and it has been reported to have a DNA matching fraction of 0.44 in its three most-populous counties in England. However, this was a pioneering result that has since been questioned. Plant has a UK population of 13,396 and a matching fraction of 0.5 in the UK. Together, they suggest values approaching a high-end limit for the size of the matching fraction in a moderately-common UK surname.

At the other extreme, Bray and Jefferson are low-matching moderately-common surnames. Bray has a UK population of 17,588 and its lack of any DNA clustering in a sample of 29 random volunteers implies that its largest fractional descent cluster must be less than 0.07 of the whole selected sample. Jefferson has a UK population of 5,906 and a K&J sample size of 85; its largest DNA fractional descent cluster is only 0.095 of its sample. The low clustering for Bray shows up as the gap in the purple crosses in Figure 1 and that for Jefferson as a dip for its purple cross.

We here consider some computer simulation results that are appropriate to Bray. This surname is widely distributed throughout England. Though K&J give its population as 12,195, its UK population is 17,588 according to the *worldnames* web-site. We have accordingly carried out simulations with 3,000 reproductively-active males and the all-England growth parameters. Figure 13 shows that a modal number of around 20 separate-origin descent families are predicted.

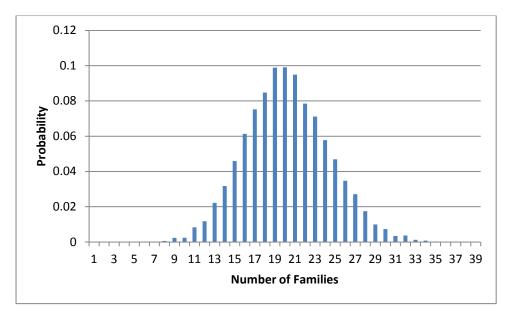


Figure 13: Predicted numbers of families in 3,000 reproductively-active males with all-England growth parameters.

Figure 14 shows predicted sizes of the largest descent family in such a surname, assuming 20 descent families. It shows that there is barely a realistic chance of the largest one being as small as 3 hundred, which is 10% of the population. For an NPE rate of 2% per generation, this would correspond to a 0.06 fractional descent cluster, below the observed limit of 0.07 in K&J's DNA sample. Thus, the simulations of Figure 14 are compatible with the observed DNA result for Bray,

provided that we consider this surname as an extreme case of little clustering. If we bear in mind however the effect illustrated above in Figure 12, the observed result for Bray (Figure 1) becomes a rather less extreme prediction of small largest family size. This effect implies that he chances of the largest predicted descent family being only a fraction 0.06 of the sample can be expected to be increased above that shown in Figure 14 if there are more than 20 separate-origin descent families. A number reaching up to as large as around 30 is allowed with some probability by Figure 13. There is also the standard error of the sample to be considered whereby a different sample of Bray men might show a larger matching fraction than 0.07.

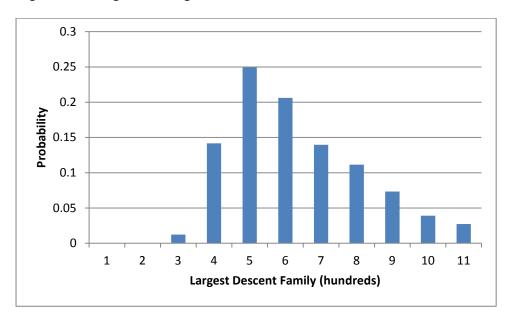


Figure 14: Predicted sizes of the largest descent family for 3,000 reproductively-active males and the all-England growth parameters, assuming 20 descent families.

# Different DNA matching fractions for Plant in the UK and in the diaspora

We can consider predictions from our computer model simulations in connection with the large observed difference between the DNA matching fractions for Plant, in the UK and overseas. For this, it is relevant to note the sample sizes of men giving rise to our observed DNA results. The limited sample sizes contribute some of the spread of uncertainty in the simulation predictions. Moreover, in the simulations, purely fortuitous effects can arise: from genetic drift; from random NPEs; from random emigration; as well as from the random sampling of DNA volunteers. These each contribute some spread to the range of expected values for the matching fractions. This is illustrated in Figure 15 by the spread of the bars. We can compare these computer predictions for a single family with our observed DNA results for Plant, for which we have the details of our sample sizes for the UK and the diaspora. We have observed a 0.5 matching fraction for a sample of sixteen Plant men currently living in the UK but 0.76 for nineteen men living in the diaspora. Such sample sizes are taken into account for our computer model simulations, along with the other random effects, in Figures 15(a) and 15(b) which both apply to a single-origin model for the Plant surname in respectively the UK and the diaspora.

If we assume an NPE rate of 1% per generation throughout (orange bars in Figure 15), there is a very small probability predicted, in a single-origin surname model, for the observed matching fraction

value of 0.5 (Figure 15(a)) but there is a reasonable predicted chance of the value 0.76 (Figure 15(b)) for the diaspora. Thus, this 1% NPE model (orange bars) adequately explains the observed DNA result for the Plant diaspora but we then need to conjecture that there is an extra source of mismatches in the UK. There is no particular reason to suppose that the NPE rate for the UK was higher than in the diaspora, so we can conjecture that the additional mismatches in the UK could have arisen from separate-origin families that are smaller than the dominant one. These small separate-origin families in the 1% NPE model (orange bars) would then explain the observed 0.5 matching fraction for the UK; and, for the diaspora, it would then seem that we would then need to assume that the smaller extra families have not emigrated as much as the dominant one.

We have noted from Figures 3 and 4 that smaller surnames seem, at face value, to have a tendency to emigrate less than more common ones. We might similarly consider that the largest family, within a plural-origin model for the Plant surname, emigrated more than those of smaller separate-origin families. Specifically, the largest Plant descent family appears to have grown abnormally in the UK and this might have produced a shortage of land for younger sons in the main Plant homeland such that this largest Plant descent family has dominated the population of the Plant emigrants. Such a plural-origin model for Plant in the UK is compatible with both the documentary evidence and with the computer predictions which allow a dominant family and several smaller ones in the UK.

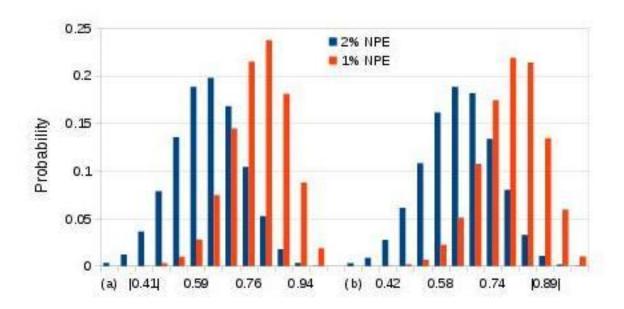


Figure 15: Predicted probabilities, in a single-origin model, of various matching fractions for: (a) a sample of 17 men in England; (b) 19 in the diaspora. The orange bars are for an assumed NPE rate of 1% per generation and the blue bars are for 2%.

It is possible however that the NPE rate was 2% per generation (blue bars in Figure 15), rather than 1%, throughout the UK and the Plant diaspora. Then both the observed matching fractions for Plant, 0.5 in the UK and 0.76 in the diaspora, have reasonable chances of arising, according to the blue-bar computations. Indeed, using more straightforward statistical calculation, the 0.5 and 0.76 values are

not significantly different at the 95% confidence level; given the sample sizes, the difference might have arisen purely from random fortuity and no further speculation is needed to explain the result. In other words, the observed DNA matching fractions do not rule out a single-origin model for Plant (blue bars in Figure 15) though it seems likely that there are some small additional descent families in the UK as outlined above in connection with the orange bars.

#### **Computer simulations for very-common surnames**

Assuming an unlimited number of possible semantic origins to a surname, arising with one or more different meanings for many fourteenth century individuals, the computed probability distribution for the numbers of UK descent families in a very common surname is as shown in Figure 16. This simulation is for a surname with twenty thousand living reproductively-active males; it assumes the all-England growth parameters. It peaks at around 130 descent families with little predicted probability of fewer than 100.

Since most descents die out, this might correspond to around a thousand medieval progenitors for the surname. However, since it is rare for UK surnames to reach the size of a very common surname, it might be that an unusually high fraction of the descent families have survived, implying rather fewer different progenitors, related or unrelated, who were ascribed the same surname. It might be questioned whether a single surname can have this many medieval progenitors, even in this exceptional case of a surname that is unusually large. As another rare event, there might be a substantial number of abnormally large families in very common surnames. This particular possibility would be consistent with an expectation that abnormally many in the surname could have experienced land shortages; that hypothesis could help to explain the relatively high emigration rates observed for very common surnames in Figures 3 and 4. Extreme behavior appears to be needed to explain very common and prolific surnames.

We are proceeding with further investigations which might shed further light. We are planning a second edition of this article.

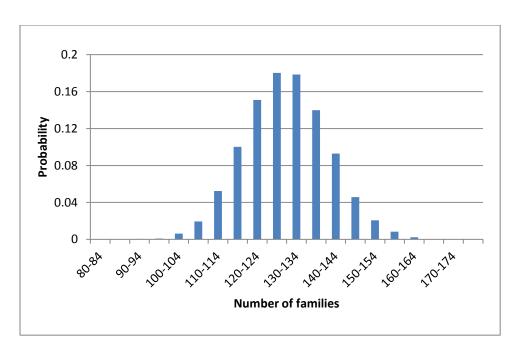


Figure 16: Predicted chances of different numbers of families in a very common UK surname

# **Appendix A: The Plant DNA results**

As far as we know, nearly all the Plants tested volunteered independently to be DNA tested. Only one was counted amongst any non-independent volunteers, typically relatives from the same family, and the rest were excluded from our statistical analysis. The resulting `worldwide' dataset for living Plants is broadly evenly distributed amongst the surname's most prevalent countries (Table 2). A chi-squared test shows that no country is significantly over-represented.

	UK	USA	Australia	Canada	Ireland	New
						Zealand
Plant population	13.3k	7.8k	3.2k	2.6k	0.3k	0.2k
In DNA dataset	16	11	6	2	1	1

Table 2: Country populations for Plant and random volunteers in the Plant worldwide sample

Fuller details of the DNA results for Plant are given elsewhere where they are being progressively updated. The full results are too extensive to include here and, accordingly, Table 3 shows just a small number of the Y-STR markers that have been measured. It corresponds to the worldwide dataset of independent volunteers, which is represented in Table 2. Some of the independent volunteers are shown in Tables 3 and 4 in red and these are the ones who are living overseas. Those living in Great Britain are shown in black. We consider general emigration rates separately from

<sup>9</sup> http://www.plant-fhg.org.uk/dna.html and http://www.familytreedna.com/public/plant/

diaspora growth rates in Appendix D where we deduce that the diaspora growth rates were apparently significantly different from those in England, in particular between the mid seventeenth and nineteenth centuries.

In Table 3, the most frequent Y-STR marker values are labelled PMH denoting the "Plant Modal Haplotype" of the matches; discrepancies from these values are highlighted in bold. Apart from PMH, a Plant project code for each independent test-taker is shown in the first column, with P denoting the spelling Plant and PT for Plantt. The column headings are the DYS labels of just the first 12 Y-STR markers only.

DYS	393	390	19	391	385a	385b	426	388	439	3891	392	38911
PMH	13	24	14	11	11	14	12	12	11	13	13	29
P1a	13	24	14	11	11	14	12	12	11	13	12	29
P1b	13	24	14	11	11	14	12	12	11	13	13	29
P1c	13	24	14	11	12	14	12	12	11	13	13	29
P2a	13	24	14	11	11	14	12	13	11	13	13	29
P5a	13	24	14	11	11	14	12	12	11	13	13	29
P6a	13	23	14	10	11	15	12	12	11	14	13	30
P7a	13	24	14	11	11	14	12	12	11	13	13	29
P7b	13	24	14	11	11	14	12	12	11	13	13	29
P8a	13	24	14	11	11	15	12	12	13	13	13	30
P9a	13	24	14	11	12	15	12	12	12	14	13	30
P11a	13	24	14	11	11	15	12	12	13	12	13	28
P12a	13	24	14	11	11	14	12	12	11	13	13	29
P13a	13	23	16	11	11	15	12	12	12	13	13	29
P14a	13	24	14	11	11	14	12	12	11	13	13	29
P16a	13	24	15	11	11	14	12	12	13	13	13	29
P17a	13	24	14	11	11	11	12	12	11	13	13	29
P18a	13	24	14	11	13	15	12	12	12	14	13	30
P19a	13	24	14	11	10	14	12	12	11	13	13	29
P20a	13	24	14	11	11	14	12	12	11	13	13	29
P21a	13	24	15	11	11	14	12	12	13	13	13	29
P22a	13	25	14	10	11	14	12	12	12	14	13	30
P23a	13	24	14	11	11	14	12	12	11	13	13	29
P24a	13	24	14	10	11	14	12	12	12	13	13	30
P25a	13	24	14	11	11	14	12	12	11	13	13	29
P26a	13	24	14	11	11	14	12	12	11	13	13	29
P27a	13	24	14	11	11	14	12	12	11	13	13	29
P28a	13	24	14	11	11	13	12	12	11	13	13	29
P29a	13	24	14	11	11	14	12	12	11	13	13	29
P30a	13	24	14	11	11	13	12	12	11	13	13	29
P31a	13	24	14	11	11	14	12	12	11	13	13	29
P32a	13	24	14	11	11	14	12	12	11	13	13	<b>2</b> 9
P33a	13	24	14	11	11	14	12	12	11	14	13	30

P34a	12	23	14	11	11	14	12	12	12	14	13	30
P35a	14	21	15	9	13	13	11	12	11	14	11	31
PT1a	13	24	14	11	11	14	12	12	11	13	13	29
PT2a	13	24	14	11	11	14	12	12	11	13	13	29
PT3a	13	24	14	11	11	14	12	12	11	13	13	29

Table 3: An extract of 12-marker Y-STR results for independent Plant volunteers.

In fact, the first 12 markers (Table 3) of the PMH (Appendix B) correspond to the WAMH3, that is the third of the four signatures that are collectively known as the Western Atlantic Modal Haplotype. Collectively, these four relatively common Y-STR signatures account for around 1.3% of all the men in Western Europe; they are particularly common Y-STR signatures within the common Y-SNP haplogroup R1. When more Y-STR markers are measured, however, the PMH becomes an essentially unique DNA signature for the main English-origin Plant family and there is then no overlap with the WAMH.

The following Table 4 outlines the quality of matching for those tested volunteers that have been designated to be a "match" to the main English Plant family. The column "GD in 12" gives the Genetic Distance of each volunteer from this family's PMH (Plant Modal Hyplotype) when only the first 12 loci are considered; the two values separated by a comma represent calculations in respectively the stepwise and infinite-alleles models. The testing company FTDNA (Family Tree DNA) uses intermediate calculations between these two bounds and the company quite obtusely changed the model used for their calculations in December 2012; nonetheless, their rather arbitrary genetic distance values remain bounded by those of our two standard types of calculation. Similarly, "GD in 25" gives our genetic distance values when only 25 loci are considered. For 37 loci and more there is a close ambiguity in the PMH as to whether the locus CDYa has the value 36 or 37 and two pairs of GDs, corresponding to the two values of CDYa, are separated by a '/'. Similarly for 67 and 111 loci.

The column "GD in most" summarises the situation for the highest number of FTDNA loci that have been measured. The column "Related to main family" gives summary conclusions based on this preceding column with such a designation "poss,prob" denoting possible or probable for this test-taker being an intact male-line descendant the main English Plant family, with two designations corresponding to the two bounds (i.e., for the stepwise and infinite-alleles models) and also, when there are four, the two CDYa modal values. The designations for "Related to main family" are taken from item 22 of the FTDNA FAQ for Understanding the Y-DNA STR results. These designations ignore other supplementary evidence such as that from control samples, Deep Clade testing, close matching to other Plants in the Y-STR cluster and any documentary evidence of genealogical kinships.

The GDs ignore the different mutation rates of different markers: FTDNA's TiP tool takes them into account. Like other aspects of more detailed investigation, the aspect of mutation rates is not taken into account in Table 4.

False positive matches to the PMH (Plant Modal Haplotype) are not impossible but very unlikely: the chance of a false positive match is 0.056 at the 12 loci level and very much lower when more loci are

http://www.isogg.org/wiki/Western\_Atlantic\_Modal\_Haplotype

tested. In other words, the PMH is not far from WAMH3 and control group considerations based on the full FTDNA dataset accordingly allow that there is a small chance that an exact match of a Plant to the PMH might be a false positive. This small chance rapidly reduces to zero at higher marker levels at which the PMH is distant from any other Y-DNA signature, including the WAMH, in the FTDNA database.

The volunteer P28b in Table 4 did not volunteer independently for testing and he is a distant cousin of P28a who has the essentially certain "related" status.

P1a, P19a and P29a have obtained Deep Clade enhancement of their "probable" (or "poss,prob/not,prob") status; this enhancement is due to their sharing a rare haplogroup despite having very different mutations from the PMH. Detailed probability calculations taking account of both Y-STR and Y-SNP findings involve the Bayesian method, as mentioned further in Appendix B.

Plant ID code	FTDNA kit code	GD in 12	GD in 25	GD in 37	GD in 67	GD in 111	GD in most loci yet measured	Related to main family	Deep Clade check
P1a	11830	1,1	1,1	3,3/4,4	3,3/4,4	6,6/7,7	6/7 in 111	probable	yes
P1b	18329	0,0	0,0				0 in 25	related	
P1c	141186	1,1	1,1	3,3/4,4			3/4 in 37	rel/prob	
P2a	277384	1,1	1,1	4,3/3,2			4,3/3,2 in 37	prob,rel/rel,rel	
P5a	11858	0,0	2,2				2 in 25	probable	
P7a	7818	0,0	0,0				0 in 25	related	
P7b	105871	0,0	0,0	4,2/5,2	4,2/5,2	7,5/8,5	7,5/8,5 in 111	prob,rel/poss,rel	
P12a	22939	0,0					0 in 12	related	
P14a	43911	0,0					0 in 12	related	
P19a	96105	1,1	4,4	6,6/5,5	6,6/5,5	7,7/6,6	7/6 in 111	probable	yes
P20a	119000	0,0	0,0	0,0/1,1			0/1 in 37	related	
P23a	144948	0,0	0,0	1,1/2,2			1/2 in 37	related	
P25a	N83079	0,0	0,0	1,1/0,0	2,2/1,1		2/1 in 67	related	
P26a	182593	0,0	1,1	1,1/2,2	2,2/3,3		2/3 in 67	related	
P28a	230023	1,1	1,1	4,3/3,3	4,3/3,3		4,3/3,3 in 67	related	

P28b	248032	1,1	1,1	4,3/3,3			4,3/3,3 in 37	prob,rel/related	
P29a	232765	0,0	1,1	6,4/7,4	6,4/7,4	10,7/11,7	10,7/11,7 in 111	poss,prob/not,prob	yes
P30a	273914	1,1	1,1	7,4/6,4	7,4/6,4		7,4/6,4 in 67	poss,rel/prob,rel	
P31a	280105	0,0					0 in 12	related	
P32a	280384	0,0					0 in 12	related	
P33a	295512	1,1	1,1	4,4/5,4			4,4/5,4 in 37	probable	
PT1a	18227	0,0	1,1	2,2/1,1	3,3/2,2		3/2 in 67	related	
PT2a	60092	0,0					0 in 12	related	
PT3a	235642	0,0	1,1	1,1/2,1	3,3/4,3		3,3/4,3 in 67	related	

Table 4: Quality of matching to the main English Plant family (note that P28b is not an independent volunteer)

The ancestral locations of both matching and mismatching Plants were shown in an earlier paper<sup>11</sup> and this is summarized for the whole of England in Figure 17. Although mismatches (blue) in the main homeland around Staffordshire (located near the geographical cluster of red towards the west midlands) might be due partly to NPEs, it is possible that the mismatches around the periphery might arise more likely from separate medieval origins.

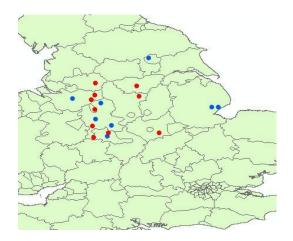


Figure 17: Ancestral locations of Plant DNA volunteers: matches (red) and mismatches (blue)

.

<sup>&</sup>lt;sup>11</sup> http://cogprints.org/9191/

# **Appendix B: Estimating the Properties of the Plant Population**

In this appendix we give a more detailed theoretical discussion of the procedures used to analyse the Plant surname. We take as a starting point of the analysis that a large fraction of the males, with the surnames Plant and Plantt, appear to be genetically related, with a most recent common ancestor who lived during the period when inherited surnames were becoming widespread in England. The genetic evidence indicates that males with the predominantly French-Canadian surname Plante are similarly related to each other but are not closely related to the Plants. In the following discussion, when we refer the Plant surname, we will mean the surnames Plant and Plantt but not Plante. To better understand the Plant surname we address the following questions:

- 1) What fraction of the Plant male population is descended from the common ancestor?
- 2) What was the DNA haplotype of this ancestor?

We will attempt to address each of these issues in as statistically rigorous a manner as possible, and for that reason some of the discussion in this Appendix is relatively technical.

#### **B.1: Estimating the fraction of genetic descendants**

In addressing the first question, we need to avoid basing the estimate of the fraction of common descendants on a biased sample. Our simulation analyses end in 2001. As discussed in Appendix D, emigrant families to the United States, Australia, and other countries grew at a different rate from those that remained in Great Britain. Therefore to avoid biasing our results with inappropriate growth conditions we will restrict our analysis to those men living in Great Britain in 2001. We are grateful to Debbie Kennett for suggesting this. Another potential source of bias may come from the use of data only from people who had already been tested and have requested a more detailed test. It is possible that individuals who are clearly not in the main family would be less likely to request further testing. For this reason in estimating the fraction of the population directly descended from the same ancestor we will restrict ourselves to the twelve loci test, which is the entry point for most of the men tested. A few had previously had only ten markers measured by Oxford Ancestors but these have almost entirely upgraded to at least twelve loci. 12

At the twelve marker level, the Plant Modal Haplotype (PMH) is given by the modal value at each locus. The PMH is not necessarily the ancestral haplotype of the first Plant to have the surname, because of subsequent genetic drift causing different family branches to grow unevenly, but it represents the best available estimate. Of the 16 men living in Great Britain, eight, or fifty percent, are within a genetic distance of one from the PMH. Of the five who are at a genetic distance of one, three have been tested at the 111 marker level and these same three have been tested for haplogroup membership. All three, including two at a relatively large 37 marker genetic distance from the PMH, test positive for membership in the same haplogroup, which is found in only six

<sup>&</sup>lt;sup>12</sup> There are just two who have not upgraded. One, a mismatch, was not a Plant but thought that he might have descended from a Plant ancestor. The other, P2b, who matched, was believed to be a distant male-line relative of P2a who matched and who has upgraded.

percent of the English population. A Bayesian analysis places the probability of these individuals belonging to the main Plant family at 98%. Therefore we will assume that all tested individuals at a genetic distance of zero or one from the PMH are in the main Plant family.

If the estimated fraction of the population who are genetic descendants of the original ancestor is p = 0.5 and n = 16 is the sample size, then the standard error is given by

$$s = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5 \times 0.5}{16}} = 0.125.$$

Therefore a 95 percent confidence interval is  $0.5 \pm 1.96 \times 0.125 = 0.5 \pm 0.245$ , which is the interval between 0.255 and 0.745.

The confidence interval is rather large because the sample size is rather small. The standard error computed above is based on the assumption of an infinite population, but the actual population is of course finite. Irvine<sup>13</sup> implicitly refers to this fact of a finite population size in his concept of penetration, which is defined as the fraction n/N, where N is the size of the total population. The UK population of males with the surname Plant is approximately 13,000. To take into account the finite population size, one multiplies the standard error by a *finite population correction* given by

$$fpc = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{13,000-16}{13,000-1}} = 0.999.$$

It is evident that the finite population size has a negligible effect on the precision of the estimate. In general, the effect of the population size is only important for very small population surnames. The primary effect on precision in our case is not the low penetration per se, but rather the small size of the sample itself.

#### **B.2: Estimating the Modal Haplotype**

We follow the normal procedure of estimating the marker values of the original Plant ancestor by the modal values of the sample. To obtain the most accurate estimate of the mode, this sample should only include genetic descendants of this ancestor. Here we run into a case of circular reasoning in that we are trying to determine the members of the sample through their genetic distance from the PMH, but the PMH is determined from the sample itself. Therefore we must proceed cautiously. We begin by again using the 12 marker data. Fig.1 shows the histogram of stepwise genetic distances from this modal value of all 37 of the Plant males tested at the twelve marker level. These values fall into two groups: those at a genetic distance of zero or one from the modal value, and those at a genetic distance greater than or equal to three. The computation of genetic distance follows the standard practice of subtracting the value of DYS385a from that of DYS385b. According to the FTDNA rules of thumb, individuals at a genetic distance of one at the

<sup>&</sup>lt;sup>13</sup> J.M. Irvine (2010) *Towards improvements in y-DNA Surname Project Administration*, Journal of Genetic Genealogy, 6(1), 1-24.

twelve marker level are "probably' related, and those at a genetic distance of three or more are not related. It therefore makes sense in determining the most likely PMH to exclude those individuals at a genetic distance of three or more from the mode. To be conservative, however, we will initially include all individuals. This does not change the value of the mode of any locus.

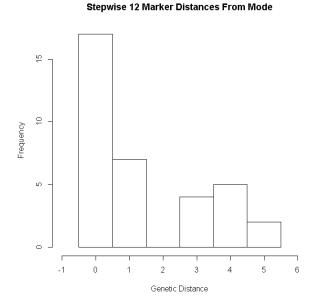


Figure 18: Histogram of genetic distance from mode.

The locus with the greatest number of individuals having a value different from the mode is DYS385; eleven of the 37 independent samples differ from this value. Intuitively, it seems clear that the chance is very slim that further sampling could reveal a PMH different from the modal values of the current sample. One way to make this statement more precise is to compute a bootstrap estimate of their standard error. The bootstrap of the standard method for estimating the standard error of those statistics whose standard error, unlike the fraction *p* above, cannot be computed according to a formula. In brief, one samples the data with replacement many times (these are called "resamples"), and computes the statistic, in this case the mode, for each of these samples. One then computes the standard error of these resamples. When this computation is carried out for the mode of the 25 samples that make up the membership of the main family, the estimated standard error is zero to 3 decimal places for all of them. This very small standard error strengthens the argument that the PMH is the set of modal values.

To repeat the analysis at the 37 marker level we again compute the PMH based only on those individuals who are within sufficient distance from each other to be considered related within surname time. Figure 19 shows the resulting histogram of stepwise genetic distances from the PMH. The subpopulation of sixteen apparently related individuals again separates itself from those that are apparently not related. Of the sixteen individuals, seven have a CDYa value of 36, seven have a

<sup>&</sup>lt;sup>14</sup> R.Y. Rubinstein (1981). Simulation and the Monte Carlo Method. Wiley, New York.

value of 37, and two have a value of 35. Therefore we conclude that the modal value of this locus cannot be determined with certainty at this time, and that calculations involving the PMH should be carried out using values of both 36 and 37 at this locus.

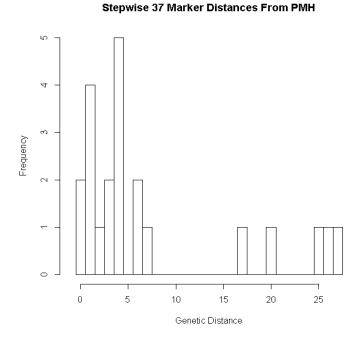


Figure 19: Histogram of stepwise genetic distances from the presumed Plant Modal Haplotype.

# Appendix C: The less certain data for Sykes

In the initial Sykes and Irven study,<sup>15</sup> there were 48 volunteers with the Sykes surname from the English counties of West Yorkshire, Lancashire and Cheshire. We shall refer to this as the "Sykes UK-3-counties" dataset. A limitation is that only four Y-STRs were measured so that it is unclear, for example, whether clusters of 2 or 3 were just fortuitous matches: these accounted for 35% of the dataset of 48 volunteers. Of the remainder, 44% were in a cluster of 21 matching volunteers and 15% in a cluster of 5.

More recent results<sup>16</sup> for the surnames Sykes and Sikes are apparently "USA-biased". Table 5 shows the predominant countries for the general population of the Sykes surname. It is uncertain that the volunteers "USA-biased" DNA dataset were random, rather than obtained by soliciting more than one volunteer from each of several particular families. Hence, clusters of 2 or 3 might not be of independent random volunteers though they accounted for 18% of the "USA-biased" dataset of 79. The 44% cluster in the "UK-3-counties" dataset became 23% in the "USA-biased" dataset and the 15% cluster became possibly more than one cluster amounting to 8%. In the "USA-biased" dataset,

\_

<sup>&</sup>lt;sup>15</sup> Bryan Sykes and Catherine Irven (2000) `Surnames and the Y chromosome', Am J Hum Genet 66(4), pp. 1417–1419.

<sup>&</sup>lt;sup>16</sup> http://sikes-sykesfamilies.rootsweb.com/dna-chart.htm

there were also clusters of 25%, 11% and 9%. One possibility is that non-random emigration from the UK and genetic drift in the USA gave rise to these different clustering characteristics in the "USA-biased" dataset though, as already mentioned, the extent to which this dataset represents random and independent volunteers is uncertain.

	USA	UK	Australia	Canada	New	Ireland
					Zealand	
Sykes	27.3k	21.4k	4.7k	3.5k	0.6k	0.1k
population						

Table 5: Predominant countries for the Sykes surname

# Appendix D: The Simulation Model

In this appendix we describe the simulation model used to explore the probabilities of outcomes associated with unusual population growth. Our objective is to estimate the probability distribution associated with a surname line begun by a single progenitor in 1311 and continuing to 2001. We address this objective using Monte Carlo simulation. In this technique, the model includes random variables whose values are generated according to a probability distribution. In our case, the random variables are the number of sons born to each family and surviving to reproduce in the next generation and, for each son, the event of emigrating from England, and the event of having a different father from that providing the family surname. The Monte Carlo simulation functions by simulating the dynamics of a single surname family a very large number of times (in our case, one hundred thousand times) and taking statistics on the outcome of the simulations. These statistics are considered to provide an accurate estimate of the distribution of possible outcomes of the system being simulated (in our case, the growth and dynamics of a surname population).

The model keeps track of reproducing males. The basic model assumes a 1:1 sex ratio and considers only males that survive to procreate in the next generation. For purposes of brevity we will not continue in the description of the model to specify that we only include male offspring who survive into adulthood; this will be implicit in the discussion. The males are divided into four categories: residents of England who are genetic descendants of the progenitor (denoted EG), residents of the diaspora who are who are genetic descendants of the progenitor (denoted DG), residents of England who are descendants of a non-paternal event (NPE) (denoted EN), and residents of the diaspora who are descendants of a non-paternal event (NPE) (denoted DN).

The model functions in discrete generations. For each individual in a given generation, the model first determines the number of sons. If the individual is an EG, the model then determines for each son (a) whether that son emigrates and becomes a DG, (b) whether that the son is an NPE and becomes an EN, and (c) whether both occur and the son becomes a DN. These determinations are made based on the value or a computer generated random variable as described in the next paragraph. Similar determinations are made of potential outcomes that move sons among other classes. All movement is unidirectional: obviously an individual cannot go from being an NPE to a genetic descendant, and rather than allowing reverse migration, the parameters of the model are adjusted to consider only net migration. Thus a son of an EG may become an EG, DG, EN, or DN; a

son of a DG may become a DG or DN a son of an EN may become an EN or DN, and a son of a DN can only be a DN.

At each generation the dynamics of the model are therefore determined by three random variables, each generated from a different probability distribution These are the probability distribution governing the number of sons born in each family, the probability that a son born to an English family will emigrate, and the probability that a son born to a family of genetic descendants of the progenitor will be an NPE. The number of sons born to each family and surviving to adulthood is assumed to be a random variable drawn from a Poisson distribution. This is the most common way of selecting the size of a group of individuals, and is used in all simulation models of this type. The Poisson distribution is characterized by a single parameter: the mean (in our case, the mean number of surviving male children in each family). The probability distributions governing emigration and an NPE are both binomial.

The parameter characterizing the Poisson distribution governing the number of sons is computed according to the theory of branching processes, as described by Pinsky and Karlin,<sup>17</sup> from the rate of population change in each generation (i.e., the intrinsic growth rate). English population data were interpolated from Broadberry et al. for the period from 1311 to 1541, from Wrigley and Schofield<sup>18</sup> for the period from 1541 to 1801, and from census records for the period from 1801 to 2001.Estimation of intrinsic growth rates for population growth in the diaspora was more difficult because the population of English descendants will have mixed with descendants of other ethnicities. Moreover, this population growth was due to a mixture of immigration and intrinsic growth. It was assumed that all diaspora population growth prior to 1791 occurred in the United States, and that no mixing occurred with other ethnicities during this period. The total number of English immigrants to the USA prior to 1791 was 230,000, about half of whom were reproductive age males.<sup>19 20</sup> The English and English descendant population in 1791 was 2,100,000. Emigration rates and initial intrinsic growth rates were computed from these data based on the assumption that English emigration was proportional to the population of England and that the intrinsic growth rate was constant.

The probability that an individual Englishman would contribute to net emigration was assumed to have a constant value of 0.05 per year between 1821 and 1911. No net emigration was assumed to occur in 1791, 1941, or 1971. Based on available census data, reproductive age males continued to constitute about half this number. Immigration rates per generation were computed based on these data. The method described in the previous paragraph to estimate the intrinsic growth rate was not used for post 1791 values because of the mixing problem described above. Instead, the growth of the English surname population in the diaspora was estimated as being proportional to the growth

<sup>18</sup> E A Wrigley and R S Schofield, *The Population History of England 1541-1871: A Reconstruction* (Cambridge University Press, 1981).

<sup>&</sup>lt;sup>17</sup> M A Pinsky and S Karlin, *An Introduction to Stochastic Modeling* (Academic Press, Boston, 2011).

<sup>19</sup> http://en.wikipedia.org/wiki/History\_of\_immigration\_to\_the\_United\_States; http://www.trivia-library.com/b/people-races-ethnicity-in-the-u-s-english-americans-part-2.htm.

<sup>&</sup>lt;sup>20</sup> Jones, M. A. The background to emigration from Great Britain in the nineteenth century. Fleming, D. and Bailyn, B., Editors. Perspectives in American History, vol. 7 pp. 3-94. Charles Warren Center for Studies in American History, Harvard University (1974).

http://en.wikipedia.org/wiki/History of immigration to the United States; http://www.trivia-library.com/b/people-races-ethnicity-in-the-u-s-english-americans-part-2.htm

of the four most common English surnames: Smith, Taylor, Brown, and Wilson. Relative intrinsic growth was computed from populations of these surnames in the 1850, 1880, 1910, and 1940 censuses. The final absolute growth rates of the diaspora were then computed based on the ratio of the total population size of individuals with these surnames in the USA, Canada, and Oceana to the size of this population in Great Britain. The growth rates used in the model for all England, for the county of Staffordshire in England, and for the diaspora are shown in Figure 20.

# - England - Staffordshire - Diaspora

England, Staff, and Diaspora Growth Rates

Figure 20: Plot of the per generation growth rates used in the model

In models that include early polygyny, we model polygyny by having each male in a generation in which polygyny is practiced generate a number of male offspring equal to the sum of n Poisson distributed random variables, where n is the number of wives of the male, and has a fixed value for each generation in which polygyny exists.

The generation time is a key variable in the simulation. Although the human generation time is often taken to be about 25 years, recent research suggests that it is longer, possibly as long as 35 years. We use a generation time of 30 years. This is based on the assumption that the generation time can be taken to be the mean maternal age at birth. Wrigley and Schofield show that in England this age had a consistent value of about 31 to 32 years from the sixteenth through the nineteenth century. We took the value of 30 as a round number that reflects a possibly shorter generation time during earlier centuries.