Journal: www.jogg.info
Originally Published: Volume 6, Number 1 (Fall 2010)
Reference Number: 61.011

Y-DNA PROJECTS: DEFINING A METHODOLOGY TO RECONSTRUCT THE FAMILY TREES OF A SURNAME WITHIN A DNA/DOCUMENTARY DUAL APPROACH PROJECT

Author(s): Chris Pomery

Defining a Methodology to Reconstruct the Family Trees of a Surname Within a DNA/Documentary Dual Approach Project

Chris Pomery

Abstract

An earlier article in JOGG made the case that a genealogical project aiming to reconstruct the family trees of a single surname, or group of related surnames, should integrate Y-chromosome DNA and documentary data in a dual approach methodology rather than relying on one set of data or the other in a single purpose project (Pomery C, 2009) (www.jogg.info/52/files/pomery.pdf). This follow-up article sets out a core methodology for conducting a whole surname reconstruction exercise utilising both sets of data, describing the phases and milestones such a project will pass as it progresses towards completion.

While the discussion that follows relies on Y-chromosome DNA results, focuses on surnames originating in England and describes a surname reconstruction project starting ab ovo, the methodology outlined and conclusions reached are broadly applicable to any type of geographical selection or surname frequency and can be adapted to expand single approach reconstruction projects of either type already underway. The use of Y-chromosome DNA testing to investigate the process evolution of surnames is outside of the scope of this paper, and developments in this new field may lead to revisions to the methodology outlined here.

Introduction

Whole surname documentary studies have existed in the UK for many years, typically being confined to surnames with a low frequency in the general population or those identified with a highly specific geographical area of origin.[1] However, the online publication of large-scale indexes and datasets such as national censuses and civil registrations has made it feasible to reconstruct the trees of relatively higher frequency surnames on a national basis. While the Guild of One-name Studies, a UK-oriented society promoting surname studies, currently has around 2,350 members, my personal estimate is that the total number of researchers worldwide currently undertaking a documentary reconstruction project for a UK-origin surname is at least three times that number. A key aim of most researchers is to track the surname back to its geographical and temporal point of origin.

The first surname-defined Y-chromosome DNA study started in 1997.[2] The connection between surnames and genetics has since attracted several academic studies (King & Jobling, 2009) and more than 5,000 surname-defined Y-chromosome genealogy projects are now underway, the vast majority hosted by Family Tree DNA based in Houston, TX. The majority of these DNA projects appear to be administered by a US-based researcher and the initial scope of the project is often to link US-resident name-bearers to a specific immigrant ancestor.

It is a common feature of both documentary and Y-chromosome DNA surname projects to include additional surnames within their remit. In the case of documentary projects these are generally thought of as lexical 'variants' of a core surname, the hypothesis being that collectively they constitute a single 'name.' Documentary-led projects tend to be parsimonious when

[1] The United Kingdom (UK) covers England, Wales, Scotland & Northern Ireland. The UK together with the Republic of Ireland is known as the British Isles. For fuller definitions see http://en.wikipedia.org/wiki/Terminology_of_the_British_Isles

Address for correspondence: Chris Pomery, DNAresearch@pobox.com

Received: May 1, 2010; Accepted: Oct. 18, 2010; Published: Dec. 19, 2010

[2] The Savin surname project run at University College, London, by Alan Savin. See: www.isogg.org/wiki/Timeline:History_of_genetic_genealogy

adding variants as each one represents a significant additional investment in time and money to research effectively. DNA-led projects are comparatively more promiscuous as each project's set of DNA results becomes more informative the greater the number of potentially useful results they are compared with. A central mechanism to assist DNA project managers to identify which additional surnames could provide the most useful comparisons has yet to be developed, whether organised on a geographical basis (i.e. surnames that appear to originate in the same place), lexical similarity (i.e. surnames that sound or look alike), or historical connection (i.e. surnames with a known or hypothesised connection).[3]

Scope of this article

Given that this paper is predominantly addressed to an audience of DNA project managers, I will structure the following discussion from the point of view of a DNA project manager running a complementary documentary project. I should state at this point that I am not setting up the methodology described here as a formal orthodoxy that I advocate must be adopted by all DNA project managers. There is plenty of scope for Y-chromosome DNA surname projects to hold varying goals, for example in terms of geographical focus, and the passive 'fishing trip' collection of DNA results adopted by many DNA project managers is certainly a viable route to take at the outset of a single purpose project. My aim here is simply to describe a methodology that allows managers to begin and then to develop a parallel documentary project aimed at reconstructing the trees of their target surname(s) in the country of origin.

Documentary-led surname project managers will recognise that the methodology outlined below is directly applicable to their projects while the lead researchers of existing dual approach surname projects will be able to identify which stage their own project has reached and how the methodology can be adapted to suit their individual purpose. Data from the ongoing Pomeroy reconstruction dual approach project has been included in order to help project managers with this process.[4]

Outline of the methodology

The basic premise of the methodology is that by documenting the trees of a surname one can transform the value of each Y-chromosome DNA result so that it is associated not just with the living individual who donated the saliva sample but to a specific historical ancestor within a documented tree. By testing a minimum of two male descendants within the same tree one can identify the Y-chromosome DNA signature of their common ancestor. Further, by reconstructing their entire same-surname tree one can identify which living males need to be DNA tested in order to confirm that the shared DNA signature was held by the oldest ancestor in their tree with whom it is possible to associate a DNA test result.

Note that the methodology described here defines a 'tree' as a documented set of same-surname relationships with living male descendants. Any surname-based documentary reconstruction project will contain unlinked individuals, couples and multi-generational groups; within the context outlined here, however, if they do not have living male descendants they are not classified as a 'tree' but simply as a set of unlinked documentary data. From the documentary perspective, reconstructing all the trees of a surname allows one to state at a finite point in the dual approach project that there are no unknown trees that remain to be DNA tested. In other words, to reach a stage where the combined DNA/documentary project has created the most detailed picture possible of the Y-chromosome DNA signatures of the oldest testable ancestors in the trees which have living surname-bearing descendants and thus collectively define the surname.

As the dual project progresses the DNA research strand will identify a number of trees which share the same DNA signature, groupings described here as 'genetic families.' The DNA evidence that links them suggests that these trees ultimately belong within a single tree, and as the documentary work progresses an increasing number of these genetic families will each coalesce into a larger, single unified tree.

The DNA project is completed when:

1. there are no trees remaining to be tested that can be tested, and
2. there are no genetic families undocumented as a single tree.

Thus a dual approach project will in time reach the point where further DNA testing will create no (or only marginal) additional value in terms of uncovering or confirming documented relationships. As this point approaches it will become increasingly obvious to the project manager that the final resolution of the question of the origin of the surname can only be provided by further documentary research. The documenting of each tree needs to be undertaken as part of a 'retreating baseline' strategy, taking each tree back generation by

[3] This may change by 2014 as the results of the Family Names of the United Kingdom project are put online.

[4] A detailed description of the structure and results of the dual documentary/DNA Pomeroy project is the subject of a future paper.

generation progressively using national datasets, parish-level records and then back even earlier through other medieval records.

It should be noted that the documentary research cannot ever be completed in the sense that every loose end in every tree is resolved, as this is unachievable. In its role as described here, to complement a DNA project, it simply needs to be complete enough to define accurately the total number of trees within the surname at key points in its history.

As the final stage of documentary research gets underway, the geographical origins of the remaining trees needs to be mapped. This exercise often makes the project manager aware of potential connections that it is easy to overlook when handling data from multiple parishes. During this stage the documentary research progresses by considering two opposing hypotheses to account for the surname's origin: first, that all of the trees share a single ancestral origin, and second that the surname originated in more than one place and time with more than one individual.

Based upon the available and broadly anecdotal evidence, my expectation is that:

1. surnames of low frequency in the present-day UK population will generally turn out to have a single originating ancestor within a genealogically relevant timeframe;
2. at greater than low frequencies a surname is relatively more likely to have more than one originating ancestor;
3. some reasons behind the formation of a surname, e.g. one named after a specific location, will more likely lead to the single ancestor pattern than others.

While King has noted that Y-chromosome samples associated with low frequency surnames are more likely to match each other than those associated with high frequency surnames (King, 2009), as yet there are no data available to help define the boundaries of 'low' or 'high' frequency surnames or to compare the process of surname formation using Y-chromosome DNA data.

Within the broad framework described above each project manager will need to review three key issues in their own project:

1. the problems associated with working with groups of variant surnames rather than reconstructing a single surname;
2. differences in the process of surname formation and transmission even in contiguous countries;
3. how many centuries define a 'genealogically relevant timeframe' for each country.

Reconstructing the trees of a group of surnames is more complicated on both micro and macro scales than doing so for a single surname. Thus while many documentary-based surname reconstruction projects report that variant spellings often appear to be geographically specific in origin, a variant spelling of a surname can often originate in more than one place and time, particularly if the reason the variant appears is the same in each case, e.g. it is spelled in error instead of the predominant form of the surname. On the macro level, a variant needs to considered as part of a combined 'name' which unites it with the dominant or host form and other variants. Looked at collectively, this 'name' may well show a consistent connection to a compact geographical area that is not as visible when looking at each constituent variant. It needs to be remembered that the written forms of a surname have been subject to a great deal of change and in many cases the final, modern form in any given tree, or family in it, may only have become fixed during the past few generations. No papers have yet been written describing spelling variation within surnames and incorporating DNA evidence.

While the discussion in this paper focuses upon research primarily based upon English records and surname development within an English context, it should be noted that the constituent parts of the British Isles each have a different history of surname formation and transmission. Thus care needs to be taken not to assume that the same premises can be held when analysing surnames of Welsh origin, where the association of a surname handed down the male line was established later than, for example, in England.

These differences also affect our definitions of the time span of the 'genealogically relevant timeframe' in each country, which is determined both by the practices of surname transmission and by the availability of documentary records used to corroborate a surname's history. A detailed consideration of the differences among the constituent areas of the UK and Ireland lies outside the scope of this paper, though for the purposes of the following discussion relating to England and English-origin surnames we can consider that surnames were generally adopted around seven centuries ago and the earliest systematically collected local documentary records no earlier than around 450 years ago.

The following discussion of a dual approach methodology divides the stages that the two strands of such a project, DNA and documentary, will go through in order to make it clear how each strand develops within

the combined project. In practice, though, individual surname projects will likely advance one strand more rapidly than the other and both strands' research may well straddle different phases and milestones. This is perfectly manageable: the division of each strand into distinct stages here is simply a device to make the progress between each clearer for the reader. That said, the documentary work is more likely to progress effectively in sequential phases whilst the DNA research can be developed with more freedom to approach its various milestones. (In the following outline, 'phase' is used to describe a phase of documentary research whilst 'milestone' refers to the stage of the DNA research.)

Documentary Surname Project Methods

The goal of a documentary-led surname project is to reconstruct all its constituent family trees back from the present day to their point of historical origin in time and space. Documentary projects tend by their very nature to seek to answer the question "where does this surname come from?" A second feature is that its focus will tend towards prioritising the histories of the male name bearers specifically because their life histories have such an impact on the transmission and development of the surname.

Note that there are significant differences in the kind of data available, and the form in which it is presented, between different parts of the British Isles, namely England & Wales, Scotland, Northern Ireland, the Channel Isles, and the Republic of Ireland. The following discussion focuses on records from England & Wales only unless otherwise specified.

Documentary Phase 1: the present-day back to 1841

The standard methodology behind a documentary reconstruction project is to work back in time, generation by generation, from the present day. Such studies often start with oral and personal sources, focusing on the researcher's immediate family, before expanding to include a range of documentary records and covering all name bearers.

An initial problem for a documentary project is that there is no single public data source that provides an accurate list of living name bearers. In the UK context, the best list to work from is the last edition of the national electoral roll (covering England, Wales, Scotland and Northern Ireland) prior to 2002, after which date it became possible for voters to exclude their details from the published version of the roll. No pre-2002 list is viewable online, though commercial versions on CD can sometimes be purchased second-hand.[5] Subsequent editions of the public and edited electoral roll, and other up-to-date sources such as the British Telecom list of landline phone numbers, perform markedly less well individually or collectively to create a present-day baseline list of name bearers.[6]

A baseline list taken or built up from any source or sources needs to be compared against the total number of records noted in the online Office of National Statistics (ONS) surnames' list, valid as of 2002, in order to check that it represents a viable estimate of the number of living name bearers in England and Wales.

In the case of England and Wales, the most important set of documentary records in any reconstruction process are the national records of births, marriages and deaths from 1837 to the present day (known as 'civil registration' records). The comprehensiveness of these records is not uniform as only in 1874 did it become a legal requirement for individuals to report events. Common sense suggests that early under-reporting was probably higher for births than for deaths or marriages, though the degree of under-reporting is still much debated.

The public index to these civil registration records has always supplied additional data that greatly supports the process of linking individual event records together to form the profile of a specific person. For birth records the maiden name of the mother is given; for marriages, the spouse's surname; and for deaths, the date of birth or the age at death. However, this additional data is not currently available across all years back to 1837, its inclusion dating respectively from 1911, 1912 and 1866.

A government-backed online index plans to expand the coverage of this additional data back to 1837, though no date has yet been announced when this enhanced index will arrive online. Independent mass indexing projects, both those done by volunteers such as FreeBMD and by commercial firms, have already created online versions of the original unenhanced government index up to 2006.

Once the civil registration data back to 1837 has been collected for a surname, a linkage process can be undertaken to link sets of records together:

- A death record can be linked to a specific birth record based upon the date of birth or age at death cited in the index entry.

[5] UK-Info Disk, published by 192.com in various annual editions.

[6] British Telecom phone numbers can be put together using localised surname searches at <bt.com>.

- A birth record can usually be linked to a marriage record in the previous generation by matching the maiden name of the mother to the surname of the female spouse given in the marriage index.
- A marriage record can be linked to a composite birth/death record, though with less confidence where the inference is made primarily through the forename match or geographical consistency in the absence of any other data. This activity can be easier to perform using twentieth century records as it has become more common for people to have multiple middle names.

A linkage process using civil registration data back to 1837 as described above is made much more secure by cross-referencing the profiles created with data taken from the eight national censuses between 1841 and 1911. These are now online in their entirety and in most instances from multiple vendors.

While recognising that there is no agreed standard for linking record data, this is not the place to discuss problems inherent in assessing and sorting documentary data or how one can objectively choose between alternative linkage options (which often can only be resolved by purchasing further data such as copies of the full original civil registration entries).

To summarise, by cross-referencing the two sets of primary national data, the civil registration and the census records, it is generally possible to recreate most of the detail in most trees within a surname back to 1841. Present-day data from partial electoral roll and telephone directories can broadly be mapped onto this historical data, though with many gaps. There will certainly be a number of unlinked records, e.g. individuals who die or marry but who appear to have no birth record, or who are born but appear to have no death record. Increasingly there are records that appear orphaned due to the inability of the indexes to map the complexities of modern society, e.g. births where the mother has not married into the surname (so no corresponding marriage record can be found), or births where the mother is a name bearer and where her child has not taken the surname of its father.

The experience of researchers within the Guild of One-name Studies shows that:

- it is technically feasible to perform the linkage process described above;
- while the linkage process works more efficiently with low and medium-frequency surnames, it is still broadly useful (though correspondingly more time-consuming to perform) for high-frequency surnames;
- the cost of corroborating the basic linkages created by using the free-to-view civil registration indexes starts as low as the subscription fee to a single online provider of UK census data.

Based upon my experience within my own reconstruction project, my estimate is that the linkage process described above is robust enough to recreate the majority of the members of the majority of trees in the period from the present day back to 1837 for all but perhaps the fifteen hundred most frequently found UK-origin surnames.[7]

The principal output from this phase of research is a list of trees (containing living male name bearers) and the linked data supporting them. There are two additional outputs from a surname linkage activity of this type that are of great use within a parallel DNA project:

1. The total number of male individuals available as potential DNA testees is already broadly grouped into a finite number of family trees with a point of origin prior to 1837 (albeit leaving a number of male individuals unlinked).
2. A list of potential emigrants from the UK during the period has been created, namely those individuals where a birth event is recorded but no death record is visible.

The above linkage process can broadly move the baseline for a surname back from the present day to the middle of the nineteenth century, a chronological distance of some 170 years or, depending on each family's history, some four to six generations. This new baseline effectively reduces the numbers of both name-bearers and couples whose ancestry is being tracked back within the surname to roughly one-third of the present-day number.

Documentary Phase 2: from 1841 back to the 1600s

With the linkage work in phase 1 completed, the documentary project will now have built up a list of the ancestral heads of its trees each of which will have two important items of associated data: the parish where the oldest male, generally the head of the household, was recorded as living in the 1841 census and the place where he was born as recorded in the 1851 census.[8]

[7] The 1,592nd surname on the ONS list, Shipley, has 5,000 references, twice the number of name bearers in the multi-surname Pomeroy project.

[8] The 1841 census simply records whether the individual was born within the county they are enumerated in; the 1851 census is the earliest which records a specific location (though this may not specify the parish).

Research in this second phase will focus on records kept within individual parishes, principally of baptisms, marriages and burials. These are available in free-to-view collections online such as the IGI and FreeReg, in individual short-run transcriptions organised by the individual parish, including those organised by the growing body of Online Parish Clerks or by the relevant county family history society as well as by commercial providers.[9] During this phase the scope of the document types investigated will widen with some classes, such as wills and land records, proving especially useful.

The earliest parish records in England generally date from 1538. As in many cases early records are now lost, so the date at which records are first available will vary from parish to parish. Some events now exist only as secondary records in a contemporary or later transcript.

The principal output from this phase, by incorporating the results of documentary research in this period into the tree research made under phase 1, is that the total number of trees in the surname will be reduced and the degree of geographical concentration of their ancestral origins within one or more definable areas will be increased.

At this stage it is beneficial to estimate the number of name-bearers and couples alive in intervals back from 1841 to the advent of parish records in order to get a feel for how completely the trees have so far been researched.

Documentary Phase 3: prior to the 1600s

A discussion of the kind of records available in the medieval period lies outside the scope of this article. Although no true nationally-collected datasets similar to the Victorian censuses exist, a number of close proxies can be found. While many have been transcribed and put online, there's presently no easy way to search them.

All but the very largest surname reconstruction projects will find, however, that during this phase they will be looking for data associated with a relatively small number of parishes and their surrounds, all of which may be closely bounded within a tight geographical area. This data is, however, hard to target as many of the key documentary sources are not organised on a parish basis.

The principal output from this research phase is to arrange all the remaining trees into clusters using the best mix of information available from all sources, i.e. historical, documentary and genetic. These tree clusters represent the dual approach project's working hypothesis of how the presently unlinked trees may turn out to be linked together.

Parallel Y-Chromosome DNA Project Methods

A surname DNA project running without a parallel documentary reconstruction activity is, broadly speaking, a net collecting the Y-chromosome DNA result of any male name bearer who wishes to pay for a DNA test. Under this approach, the DNA results collected within the project will be biased in two ways relating to the trees within the surname under study: firstly, towards residents of countries that are more receptive to the benefits of DNA testing (principally the USA), and secondly towards members of already documented trees. The latter observation may seem counter-intuitive; after all, one might expect men who've done no family history research to realise that they have the most to gain by taking a DNA test. In practice, it seems that those men whose family members have already done some research, or who have some inkling that they belong to a particular tree, may well be the first to pay for a test to try to confirm it, thus weighting the results for the surname as a whole towards those trees that have already been partially documented.

A third source of bias is that two different populations, for example England and the USA, will not have faced the same conditions in their recent history and it is thus likely that name bearers in different countries will have reproduced at different rates. This factor also seems to be occur within same country populations as a recent study of 40 British surnames reported that variation in reproductive success was a key factor in the degree of genetic diversity found within the surnames they studied (Jobling & King, 2009). No surname DNA project has formally reported on the different rates of reproduction in different populations revealed in its test data, but anecdotal evidence from several projects confirms both its presence and influence within surname-led Y-chromosome projects. This is important as any use of the project's DNA results to identify the dominant haplotype within the surname, or the distribution of haplotypes within it, must use results taken from testees resident within the population of origin, not those linked by an emigrant ancestor to another population. Clearly the impact of this aspect of the founder effect is less visible in recent emigrations, but its distorting effect is relatively stronger in families linked to emigrant ancestors in the 1700s and earlier, a scenario typical of many American trees.

The creation of a parallel documentary reconstruction activity removes the effect of all three of these identified biases by associating each Y-chromosome DNA result not solely with a specific ancestor but with a specific tree

[9] Many transcriptions are now hosted online at GENUKI, accessible via the relevant parish page under each county.

within a defined set of trees comprising the surname. The documentary project sets a boundary for dual approach surname projects by quantifying an upper limit for the number of trees within the surname and creating estimates for the number of members, historical and presently living, in each tree. It also defines those trees appearing to originate outside of the country of origin and allows the analytical process to focus clearly on DNA results taken from within the surname-origin population.

DNA Project Milestone 1: one test result per tree

A DNA project incorporating information created by a parallel documentary project is set up to take a targeted approach and systematically test one male from each documented tree. The results of such an exercise will often rapidly reveal a series of 'genetic families', i.e. it will identify a set of individuals with identical (or closely matching haplotypes) found in a number of unlinked trees all of which can be hypothesised on the basis of the DNA evidence to belong to a single as-yet-undocumented tree.

Note that while some trees appear to have no living descendants today resident in the country of origin, they may hold records of men who emigrated at a specific point in history and who today have descendants living outside the country of surname origin. In the case of trees where, because there are no living descendants in the country of origin, the men being tested live outside it, the oldest potentially verifiable genetic ancestor in the tree will be that original emigrant. While his DNA signature may not provide genetic proof of this man's link to a specific tree originating in the home country, in many cases it will provide a link to a group of trees within a recognised genetic family.

The DNA results associated with each UK-origin tree collectively create a matrix of DNA signatures that any emigrant's descendant anywhere in the world can use to identify the correct tree or genetic family that they belong to in the UK.

The principal output created by milestone 1 in the dual approach project is the identification of the genetic families where targeted documentary research is most likely to be able to reduce the number of trees in the project through documented combination.

DNA Project Milestone 2: two test results per tree

The second milestone is to test a minimum of two men in each documented tree. This DNA cross-referencing removes the possibility that a single DNA result linked to any tree does not, in fact, reveal the haplotype of the entire tree but merely of the tested individual's personal line which potentially could have been replaced by external DNA at any point. With two or more men tested per tree, the shared DNA signature identified is not then associated with the individuals who have been tested but with the specific male ancestor they share in common in their tree. At this point, one can say that the DNA signature belongs to an historical figure who is located both in time and space.

In cases where the two DNA results are not the same the project leader will firstly investigate the documentary evidence to re-verify its accuracy and the connections built using it, and secondly to identify a third DNA test participant sharing the oldest common ancestor with the initial pair of testees.

The principal output by milestone 2 is a refined list of the constituent family trees in the surname showing whether the associated DNA signature is part of a genetic family or found only once in the project.

DNA Project Milestone 3: two or more test results per tree linked to the oldest testable ancestor

The third milestone extends the testing process to ensure that the DNA signature associated with the tree is that of the oldest male ancestor in the tree whose DNA signature can be verified., i.e. the ancestral DNA result associated with the tree.

Note that in some instances the DNA testing programme will need to be restarted and extended to cover those cases where documentary research reveals that the currently tested ancestor is a descendant of an earlier illegitimacy within the tree. This fact may only come to light some years after initial DNA results have confirmed a unique haplotype for the tree, the new research suggesting that the true DNA signature for ancestors older than the now documented illegitimacy has still to be identified.

The principal output at this stage of the project is an assessment of how many trees have reached the point where further DNA testing will create no additional informational value for the tree reconstruction process.

The other activity identified with this stage of the DNA strand, identifying other surnames to compare Y-chromosome results with in order to uncover presently unrealised relationships, is hard to pursue as there is currently no data available to advise project managers about which are the best options per surname, other than the obvious variants that are relatively easy to identify. Surname researchers wishing to explore this avenue can review George Redmonds' work in the field (Redmonds,

1997) and undertake extensive parish register research looking out for surnames disappearing from parish registers or entering them, which might pinpoint a progressive transformation of one surname to another, or for references to name bearers holding an alias, which can indicate a switch of surname in a particular individual or family (sometimes taking place over more than a generation).

(Some dual approach project managers may at this point set an additional DNA-related goal: to identify the surname of each man whose DNA has been introduced into the trees associated with unique DNA signatures. While interesting, I consider this to be outside the scope of the current paper).

Combining data from DNA & documentary projects

Running an active documentary reconstruction project in parallel with a DNA project will lead over time to the gradual reduction of the number of trees within the combined project. This reduction process will happen most quickly where there is a genuinely iterative research process underway that combines new inputs from both the DNA and documentary sources as they arise. Thus when two men are found to have the same DNA signature, this directs the documentary research activity towards finding the common ancestor signalled by their DNA results. In the earliest stages of a dual approach project, this ancestor will often have been born since 1841 or a generation or two prior to it.

A regular pattern quickly emerges within a combined project: over time several small trees will coalesce into a single larger tree, which by then has several DNA tests associated with it through its living descendants, all of whom report a consistently held Y-chromosome DNA signature now linked to the oldest shared ancestor within the enlarged and documented tree.

Any wide scale DNA testing programme will throw up two types of inconsistencies that have to be explained, namely the presence of a DNA signature:

1. within a tree which is different from the DNA signature consistently found among other descendants in the rest of that specific tree; and
2. a DNA signature associated with the head of a tree that is different from those of other trees originating geographically close to it.

As a project progresses it will become increasingly common that the descendants of two different trees that appear to originate in the same tight geographical area return markedly different DNA results. The question then posed is which scenario is more likely: could they eventually be documented within the same tree but one of them hold a different DNA signature due to an earlier non-paternity event, or do the different DNA results point towards them being two long-established trees of different and unrelated origins? (In this context, non-paternity event includes sexual means of introducing different Y-chromosome DNA into a surname line, e.g. marital infidelity, as well as social means of achieving the same genetic end, e.g. re-adoption of the surname along the female line). One way to hypothesise an answer to this question is to look at the results linked to other trees originating in the same geographical area: do they show a dominant DNA signature or not? The confirmation of any hypothesis, however, rests on recreating the documented linkages back through further generations.

A key point to note is that as a DNA testing programme develops, the need to use contextual data to help refine the hypotheses associated with its results increases. In summary: the DNA results point to potential linkages, the contextual evidence either supports or conflicts with the hypotheses about the origin of the trees and surname as currently held, but ultimately only the documentary evidence will demonstrate how the surname's trees are actually put together.

Under the method described above, even though the DNA results are interpreted as those of the ancestral heads of trees rather than the living sample providers, one still cannot assert that the modal haplotype – the most frequently found DNA result – reveals the DNA signature of the surname's original founder. The historical number of name bearers in each tree, rather than the number alive today, is useful at this point, but it does not signal a sure answer. Some trees it turns out are old and thin, i.e. they have very few living descendants but very old origins, while others may be fat and young, i.e. they have a large number of living descendants within a tree that has grown rapidly during the most recent couple of generations. There is presently no published data on the differing levels of reproductive success one might expect to find associated with different families within UK origin surnames.

Methodological Issues

Fourteen different methodological issues have been identified which dual approach reconstruction projects will need to tackle as their two parallel research strands develop. These are laid out in the table below and described in the paragraphs following.

	In the Documentary Project	**In the DNA Project**
Project Outset	1. Choose variations of research approach based upon surname frequency 2. Adapt activities based upon the surname's country of origin	3. Expand the DNA project's focus from a single ancestor search to a global whole surname project
End Documentary Phase 1 / DNA Milestone 1	4. Complete emigrant research 5. Define the acceptable percentage of unresolved records 6. Absorb all tree corrections	7. Standardise tests within the project using 37 markers
End Documentary Phase 2 / DNA Milestone 2	8. Identify the earliest record found in each parish or emigrant country 9. Map parish origins 10. Calculate the historical frequencies of surname occurrences 11. Absorb all tree corrections	12. Test emigrants' descendants
End Documentary Phase 3 / DNA Milestone 3	13.Define the contextual data being used 14.Arrange the remaining trees into clusters	15.Expand the pool of surnames for DNA comparisons

Table 1: Timing of Methodological Issues

Project Outset

1. Choose variations of approach based upon surname frequency

Low frequency surnames can readily be reconstructed using the dual approach method. By the close of phase 1 it is often clear that the geographical origins of all of the constituent trees originate in a very tight geographical area.[10] However, a modified research approach will need to be adopted when the surname being studied is found with a high frequency as the greater the number of name bearers the more difficult it is to link documented events to the right individual correctly. One way to overcome this restriction is to allow a more permissive approach in the linkage process (i.e. one may accept a weaker 'balance of probability' argument when associating an event to a particular person rather than requiring a match across different data points, e.g. name, location, date, co-signatories). Such changes need to be approached with caution, however, as a more permissive attitude may only create the appearance of an advantage. Errors created during the linkage process will, in due course, be signalled by the DNA & documentary evidence the project collects and will ultimately still need to be resolved.

With so few results available from published or completed surname reconstruction projects, whether DNA-driven or combining documentary research, it is problematic to make more than an educated guess about the origin pattern of names at different levels of frequency. This is made even more difficult as it is already clear that some surnames are, in fact, groups of surnames. Experience within the Pomeroy project shows that the same DNA signature is often associated with a range of modern name bearers holding different surname spellings even where they are documented within the same tree. In this case, while each individual spelling of the surname appears to have multiple DNA signatures, when pooled collectively not only does the frequency of the combined 'surname' increase but the pattern of DNA signatures becomes more pronounced.

As a starting point, and open to revision as useful data becomes available, my own rule of thumb on the

[10] An excellent study to review is JohnCreer 's project on the surname Creer. Details are viewable at www.creer.co.uk and <www.ballacreer.com>.

frequency issue, relating to a surname originating in England, is:

- a low frequency surname is one recorded less than 1,000 times in the ONS list, and
- a high frequency surname is one recorded more than 5,000 times in the ONS list.

As a working hypothesis I suspect that surnames originating in England and Wales with fewer than one thousand present-day name-bearers will turn out to be more likely to have a single genetic origin than not.

Quite how many low-, medium- or high-frequency surnames have a single genetic origin is unknown. An early paper on surnames and genetics suggested that the Sykes surname, recorded has held by 19,036 individuals in the ONS list, has a single genetic origin (Sykes B & Irven C, 2000), but this conclusion was not backed by any documentary evidence and is open to methodological criticism. A more recent paper confirmed that sharing a surname significantly elevates the probability that men share a Y-chromosomal haplotype and noted, without quantifying it, that this probability increases as surname frequency decreases (King et al, 2006).

2. Adapt activities based upon the surname's country of origin

The surname featured in this article – Pomeroy – is of Norman origin and its history within a genealogically relevant timeframe is associated with England. While the method outlined here is designed to work with English and Welsh records, there are differences in the type of documentary record that is available in other parts of the British Isles which mean that the phase 1 linkage process will need to be adapted for use with surnames that originate in those parts. All but the rarest surnames will also, at some point, likely need to consider records from all the constituent parts of the British Isles in order to build up a composite picture of their distribution within it and to tie up documentary loose ends.

3. Expand the DNA project's focus from a single ancestor search

Many surname DNA projects will be expanding from an initial life as a single ancestor project, often focused on a key emigrant to the USA. Care needs to be taken to change all the literature and descriptions of the project, including on the host company's project webpage, to reflect and emphasise the global nature of the expanded project. Approaches to potential testees can now be prioritised in different ways. For example, one might prioritise those that live close to the hypothesised geographical origin or are linked to trees thought to be associated with different emigrant ancestors.

Documentary Phase 1

4. Complete emigrant research

Emigrant research will need to be undertaken relating to the major beneficiary countries from the UK, namely the USA & Canada, Australia & New Zealand, South Africa, as well other places such as India, and Hong Kong. In some cases surnames originating in specific areas in the UK may be tied to specific historical emigrations, e.g. Cornish surnames will feature prominently in the nineteenth-century emigration of miners to Chile and Mexico. In other cases the emigration to a remote or unlikely destination may be triggered by universally applicable motives, e.g. missionary work.

The methodology described here can be applied within each of these emigrant-receiving countries towards the same end: identifying all the trees associated with the surname that have living descendants.

5. Define the acceptable percentage of unresolved records

Each project will need to set its own standards regarding the percentage of unresolved documentary event records it will accept within its project during phase 1. All projects however rigorously they are pursued will inevitably end up with unlinked records: the question for each project is, what percentage is acceptable? The key point here is that the percentage needs to be low enough to ensure that no trees remain undiscovered. It matters less that each tree has unanswered questions within it, but more that all trees have been identified.

A similar assessment needs to be made of the unlinked fragments of trees (the unconnected 'twigs' well known in genealogical research), though as these by definition have no known living members their unresolved presence shouldn't alter the parameters of the overall reconstruction project.

6. Absorb all tree corrections

Most documentary or DNA projects do not start with a blank sheet of paper; instead, many will have collected trees submitted by different researchers which were put together at different times. In some cases an original piece of research will have been passed on and copied by later researchers, some of whom may have amended parts of it or added their own personal data. And in many cases, this research will be incorrect. Such a problem may be indicated by the DNA results, or it may become

apparent during the documentary reconstruction process. Even though it may not be always possible to persuade the original researcher that they have made a mistake, or to remove the faulty research from the web, the project as a whole needs to be carefully documented so that its findings can subsequently be checked or recreated and a definitive history published.

DNA Milestone 1

7. Standardise tests using 37 markers

To make haplotype comparisons with confidence it is important to compare DNA results on a like by like basis by standardising the minimum number of markers your project will work with, preferably whilst leaving open an option to extend individual testees' results if there is any difficulty in differentiating between them. In the context of a project hosted at Family Tree DNA, for example, this would set the standard test in a project at 37 markers while leaving open the option to extend to 67 markers or use a deep clade test where further differentiation is required.

Documentary Phase 2

8. Identify the earliest record found in each parish or emigrant country

An early goal during phase 2 is to identify the earliest record found for the surname in each location, whether at parish level in the UK or national level outside of it. This will often reveal early presences in locations previously not recognised and change the overall picture of the combined project.

9. Map parish origins

The close of phase 1 is a good moment to map the parishes where trees originate in order to highlight their geographical distribution.[11] This map will become increasingly useful by the end of phase 2 when one is attempting to make links between families in relatively few parishes.

10. Absorb all tree corrections

Repeating the lesson of item #6 above, bearing in mind that the documentary evidence to link individuals within trees pre-1841 is scarcer on the ground than in later years and therefore that earlier researchers are more likely to have made educated errors and false assumptions in their pre-1841 linkages.

11. Calculate the historical frequencies of surname occurrences

To get a view of how many families the project will be researching in phase 3, it is useful to calculate an estimate of the historical frequency of the project surname back from 1841 into the medieval period. One way to do this is to work out the average number of name bearers in the general population across the censuses between 1901 and 1841 and then use that average to calculate a total population of name bearers at different times. In the Pomeroy project, the number of name bearers in the 1550s is reckoned to be less than one-tenth of the level at the start of the current millennium.

DNA Milestone 2

12. Test emigrants' descendants

As the documentary work progresses it will become apparent that some trees do not appear to have descendants alive in the UK today. However, some of them will have emigrants visible within them, or potential emigrants. The latter could be individuals or family groups not found in successive censuses, or in rarer cases several siblings. Early in phase 2 all of these potential emigrants need to be followed up and their descendants identified. As the phase progresses it will become clear that the only option to identify the DNA signature of some trees is to test emigrant descendants. In practice, small-scale reconstructions need to be undertaken for each emigrant-receiving country to ensure that each emigrant has no unknown descendants.

The total number of trees in the combined project will start to rise again as more emigrant research is followed up and as more trees are brought into the project because they can potentially be DNA tested, but their number will later dip again as the DNA results suggest genetic family connections which are subsequently documented.

Documentary Phase 3

13. Define the contextual data being used

As a project progresses the primary contextual data being used is likely to be geographical proximity or connection, e.g. recognising that it was more common for residents of a particular rural parish to move towards market town A or port B than city C. Alternatively, contextual data might well be specific to a single project, e.g. where it is noted that a lawyer in a rural county had dealings with two presently unlinked trees, sponsoring

[11] A good place to learn about mapping for genealogists is Howard Mathieson's site at <http://members.shaw.ca/geogenealogy>.

the conclusion that they are in fact related to each other. Both external and internal data needs to be noted. This may in time include comparisons with other surname projects, though at present few data are available to do so.

14. Arrange the remaining trees into clusters

At this point in the combined project the key activity is to define a hypothesis for the origin of the surname. While the number of distinct trees has been reduced during the project, a number will remain. The question going forward is: how are these trees linked together? In most cases the DNA project will have identified the haplotype of the oldest shared and testable ancestor of each tree, and the number of genetic families in the project will have been reduced to close to zero (in other words, each DNA signature will be associated with only one documented tree). Furthermore, the map showing the origins of extant trees will likely show one or more geographically contiguous groups of trees.

At this point, pulling all the DNA and documentary data together, and using whatever historical and contextual knowledge you can access, the next stage of analysis is to cluster trees together into what you estimate are the most likely connected groups, i.e. into groups of trees that you hypothesise will be revealed after further research to be a single tree. These clusters then form the framework of your hypothesis against which you can test any new piece of evidence from whatever source in order to adapt or retain it.

DNA Milestone 3

15. Expand the pool of surname results for DNA comparisons

DNA projects have a great deal of flexibility to expand the number of results they compare with. Doing so can identify variant spellings not included in the original documentary research programme which the DNA evidence can suggest may be part of a single surname. There is some evidence that low frequency surnames, previously thought to be unique in origin, may often be genetically linked to other higher frequency surnames, in effect creating a 'super surname' genetically proven to have multiple previously unrecognised variants.[12]

Data from the Pomeroy Reconstruction Project

Data from this dual approach surname project is included in this discussion to illustrate how the described methodology has been used in this one instance. Readers should bear in mind that differences in surname frequency, origin and project start condition will likely lead other project leaders to adapt it to suit their surnames and particular research conditions.

The Pomeroy surname reconstruction project has been underway for more than a decade. Initially focussing on documentary research, since 2000 it has incorporated DNA data within a dual approach project. At that point a decade ago several lines had been researched back to the 1600s, but most of the post-1837 civil registration records remained unlinked. While the few researched trees were known to have sponsored emigrants to different countries, the overseas picture was largely unknown and dominated by one huge tree extensively documented a century ago that stems from a single emigrant ancestor to the USA in the 1630s.

The project aggregates data for eight distinct surnames, the principal ones being Pomeroy, Pomroy, Pomery and Pummery. Other variants found today outside the UK are de Pomeroy & Pummeroy (Australia) and Pumroy (USA) as well as de la Pomerai within it.

Measured against the methodology outlined in this paper, the Pomeroy reconstruction project today is working its way through phase 2 of the documentary research and is close to milestone 3 in the DNA strand.

Documentary Phase 1: present-day back to 1841

My estimate is that the current population of name bearers in England, Wales, Scotland & Northern Ireland is of the order of 2,500 individuals. This estimate is built up by taking the average number of adults found in the four electoral rolls in the period 1998-2002 and adding the number of birth records in the period 1985-2002 in England & Wales to account for the non-voting population at the latter date. (The present population in Scotland & Northern Ireland is relatively very small compared to England & Wales).

Of these 2,500 name bearers, I estimate that around 1,200 are male of whom around 450 are married (in this context 'married' signifies an adult partnership, i.e. it includes cohabiting as well as married pairs). The latter estimates are made using historical data aggregating seven nineteenth-century censuses which found on average that 48% of name bearers were male and around 35% of them married (i.e. excluding widowers, those enumerated as unmarried, and males under marriageable age).

[12] The only example I know of is described in Susan Meates' article in the Journal of One-name Studies, *DNA testing of tremendous value in sorting out variants in my one-name study (Part 2)*, JOONS 9(2):6-9, which is viewable at <www.one-name.org/journal/pdfs/vol9-2.pdf>.

This baseline figure is broadly corroborated by figures taken from the Office of National Statistics using 2002 data which show a total of 2,338 name bearers.

By mid-2001 we were able to identify a postal address and telephone number for 798 adult males. Using the estimate of 1,200 males in total, the remaining 400 include roughly 300 minors plus those whose present whereabouts are not visible in any of the modern sources (perhaps 100 out of a total of 140 UK-born male individuals currently documentarily incomplete or unaccounted for in the period since 1912).

That present-day figure of 2,500 name bearers is underpinned by data from two key historical datasets for England and Wales: 14,661 civil registration events (in the period 1837-2008) and census records (in the period 1841-1901) currently linked to 8,037 historical name bearers.

At the outset of phase 1, and utilising the information supplied by trees donated by earlier researchers, the 450 or so modern-day married males were linked into 326 trees, many of them single couples or small two-generation families. We knew a bit about a few trees, but virtually nothing about the great majority of them.

During the course of phase 1, as a result of the linking of civil registration and census data, that start figure of 326 trees has today been reduced to 57 trees, five of which originate in Ireland and two from early twentieth-century Russian immigrants. Additionally the project is tracking nine trees with currently documented origins outside the UK and Ireland of which six originate in the USA and three in Canada, as laid out in Table 2.

The close of phase 1 is recommended as a good time to analyse the level of unlinked event data in the project. Clearly for low frequency surnames a higher percentage of unlinked records might be acceptable compared to a very high frequency surname, but no rules of thumb yet exist. The current position in the Pomeroy project — a medium-frequency surname — shows that 88.3% of civil registration records are linked to a tree included in the DNA project. At this point I am confident that no undiscovered trees originating post-1841 and built upon English and Welsh records exist, and I suspect that was also the case when the level of unlinked records was higher than it is today, perhaps as high as 15-18%. Based upon this experience, I'd suggest an upper level of 20% within a systematic surname reconstruction project.

To sum up, the documentary work during phase 1 is critically important to reduce the overall reconstruction project to a more manageable size. By tracing all the trees back from the modern era to 1841 the project has:

- reduced the scale of the research task required during phase 2 to around one-third of that undertaken in phase 1;
- linked the 450 couples into 57 trees; and
- showed that no trees with living descendants in England & Wales, and no documented groups originating after 1841 without living descendants, remain to be discovered.

Tree Type	**All**	DNA Tested	Not Tested
UK origin (England, Wales, Scotland, NI & Channel Isles)	**50**	40	10
Irish origin	**5**	4	1
Russian origin	**2**	0	2
US origin	**6**	5	1
Canadian origin	**3**	2	1
TOTALS:	**66**	51	15

Table 2: Current DNA project tree status

Milestone 1: one test result per tree

Milestone 1 marks the initial stage of the DNA testing project. The goal at this juncture was to DNA test one man per tree. However, it not always strictly necessary to DNA test more than a single representative of a

particular tree and it is a matter of individual judgement how the DNA strand of research should be prioritised. Table 3 below shows the testing status of the 66 trees within the global Pomeroy surname project shown in Table 2.

Many of the untested trees are relatively small and some of the trees DNA tested only once have returned a result signalling their inclusion in known 'genetic families', i.e. there is less need to test a second member of this tree as a potentially documentable link to other trees has already been suggested by the initial test result.

Reviewing the status of emigrants within the dual approach project, current figures reveal that 39 of 55 UK-origin trees — roughly three-quarters — contain emigrants, and the vast majority of these have produced descendants currently living outside the UK.

Documentary Phase 2: from 1841 back to the 1600s

With phase 1 reducing the scope of the required research in the UK to about one-third of its present number, phase 2 is the right time to create some estimates of the frequency of the surname in its earlier history. Table 4 outlines some aggregated figures for the Pomeroy group of surnames using one such method. Based on actual figures from the censuses for 1851 and 1901, the average figures for the percentage of males and of married males among living name bearers is calculated forwards to the present and then backwards in time over three centuries to create a range of population figures. (Note that these figures assume the ratios between the sub-groups are constant across the centuries and are intended to suggest a median figure in a range of possibilities rather than a specific number.)

# Trees:	Not yet DNA tested	With a Single DNA result	With Multiple DNA results	**All Cases**
UK & Ireland	11	22	22	**55**
Non-UKI origin	4	3	4	**11**
Global	**15**	**25**	**26**	**66**

Table 3: Tree Testing Progress

Year	Population Estimate	All Name Bearers	*of which:* Males	*of which:* Married Males
1545	* 3,470,000	* 180	87	32
1600	* 4,811,718	* 245	119	44
1700	* 6,045,008	* 310	149	55
1801	* 8,889,674	* 450	216	80
1851	17,925,404	860	389	146
1901	32,527,843	1,543	746	280
2002	60,000,000	* 2,500	1,200	450

Table 4: Historical Surname Frequency Estimates

*Note: * estimated figures*
Source: GenDocs (1545-1801); census tables (1851-1901); self (2002)

The figures in Table 4 are ballpark estimates that the project manager can use to sense the scale of the surname as a whole. A striking feature is that the total number of males in the mid-1500s is less than one-tenth of the modern-day figure, a scale of reduction that turns most modern-day surnames into medieval rarities.

DNA Milestone 2: two test results per tree

As the number of trees with two or more associated DNA test results within the project rises, the distribution of the trees among the different genetic families can be made with increasing confidence.

Table 5 below shows how the 44 UK & Ireland origin trees that have one or more DNA results associated with them (see Table 3) divide into genetic families. As has been apparent since the early days of the DNA project, two strong DNA signatures stand out. Together they are linked to almost a quarter of the remaining trees which collectively contain not far short of two-fifths of all the historical name bearers currently linked tof trees originating in the UK & Ireland.

It is worth restating that a genetic family is simply a presently unresolved documentary problem. The goal is to document each genetic family back to a single ancestor at the head of a single tree, at which point each tree will be headed by a unique DNA signature and this table will disappear into thin air.

While the table accentuates the size of the leading genetic families, several of the documented trees associated with a unique DNA signature link together a large percentage of the total known historical name bearers: the largest of the twenty trees in this state contains 1,117 individuals while the second and third largest contain 815 and 609 people respectively. Looked at another way, almost half of all historical name bearers belong in just three potential trees: the two largest genetic families and the largest DNA-unique tree.

Approaching milestone 2 the 55 trees in the project are boosted by the inclusion of eight more where the only potential DNA testee is an emigrant's descendant.

During this stage of a dual approach project the consolidation among the remaining trees is still increasing. Here the estimated 2,500 UK-based modern-day name bearers appear to belong in trees represented by just 25 unique DNA signatures.

Milestone 3: two or more test results per tree linked to the oldest testable ancestor

It is an important measure of the completeness of the DNA strand of a dual approach project that the larger trees have DNA tested enough members to ascertain the DNA signature of the oldest male ancestor in the tree that it is possible to measure. To explain, let's take the case of a hypothetical tree which has three consistent

UK & Irish-origin trees only	# Men Tested	# Distinct Tested Trees	# people documented in the trees in this 'Genetic Family'	Percentage of all tested trees
Genetic Family 'A'	11	6	2,404	19.0 %
Genetic Family 'B'	20	9	2,361	18.6 %
Genetic Family 'C'	8	3	855	6.7 %
Genetic Family 'D'	3	2	530	4.2 %
Genetic Family 'E'	9	4	1,042	8.2 %
Unique DNA Signatures	38	20	5,491	43.3 %
	89	**44**	**12,683**	**100.0 %**

Table 5: Genetic Families of the UK & Ireland-origin Trees at mid-2010

DNA results associated with it such that we can say they share a DNA signature. The oldest shared ancestor these three men hold in common was, let's say, born in 1705 while the tree as documented goes back another three generations, say to 1625. The man born in 1705 is then the oldest ancestor whom it is possible to DNA test, and the three previous generations illustrate the process that will be undertaken in phase 3: documenting the trees

back to their origins. There is thus a limit to how far back in time within any tree that a set of DNA results can aid its reconstruction, beyond which only traditional research can venture.

Table 6 shows the current ancestor status of the 26 trees that have had more than one member DNA tested (see Table 3).

In many cases there is a multi-generational gap between the oldest documented ancestor in a tree and the oldest ancestor with whom it is possible to associate a shared DNA haplotype.

The 5 results under the category 'Another Ancestor' include:

- one tree where there are currently two different DNA signatures without it being clear which one, if either, is the true DNA signature for the tree as a whole
- two trees where the common genetic ancestor is now thought to stem from an illegitimacy, re-opening the question as to the true DNA signature for the tree as a whole
- two trees where a common ancestor has been identified but where an additional testee would be required to identify the DNA signature of the oldest testable ancestor.

Documentary Phase 3: prior to the early 1600s

At this stage of a dual approach project further research is going to rely heavily upon the detailed analysis of pre-parish record documentary documents. If it has not been created earlier, a geographical map showing the origins of the remaining trees is a re-requisite for starting phase 3.

Map 1 below shows the origins of trees in Cornwall, Devon, Dorset and south Somerset within the Pomeroy project. (For simplicity, a few outliers have been left off the map in Ireland, London, Portsmouth and Bristol). The markers are colour coded to show the five genetic families identified at this stage; grey markers indicate trees holding unique DNA signatures within the project, while white markers indicate trees that have not been, or cannot be, DNA tested.

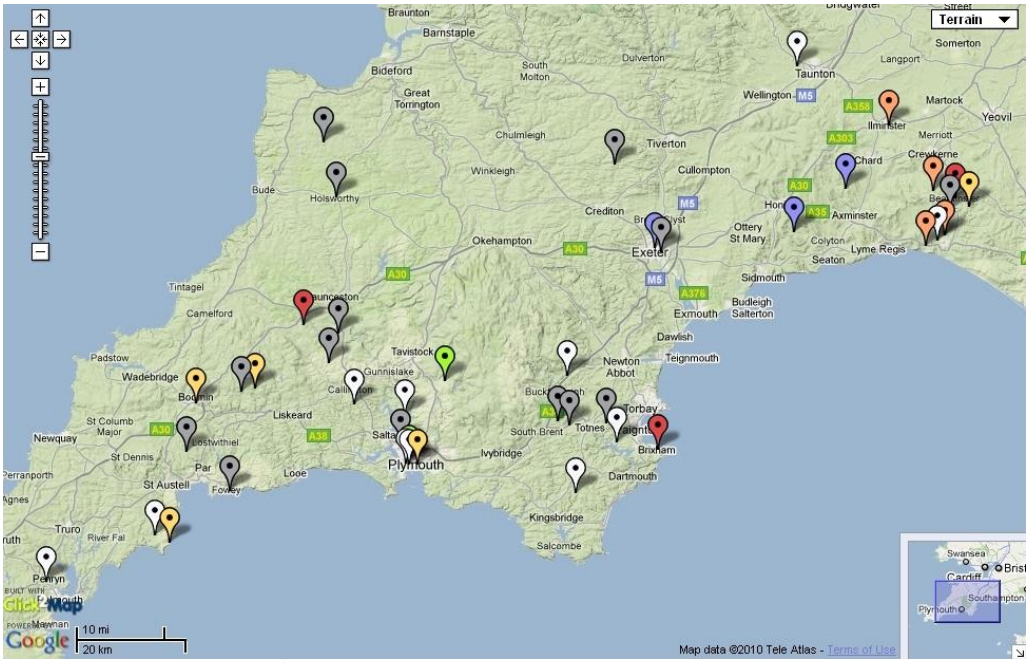
The key point that stands out is that three of the five coloured marker genetic families are concentrated in a single defined geographically area.

Thus the three blue markers (genetic family D in Table 5) originate within a few miles of each other, as do the two green ones (genetic family C) though the second of the pair is hidden among several others in the Plymouth area. Similarly the four orange markers in Dorset (genetic family E) have their origins within a few miles of each other. One can easily hypothesise how each of these genetically-defined groups will in time be documented into a single tree.

The light yellow markers in Cornwall (genetic family A) also group together, with the exception of a single tree originating in Dorset. This odd one stands out: at this stage it could be suggested that there is either a mistake in the documenting of the tree or that there was more

DNA result linked to the:	UK & Ireland	Overseas	All
Oldest Documented Ancestor	6	3	9
Oldest Testable Ancestor	12	0	12
Another Ancestor	4	1	5
All:	**22**	**4**	**26**

Table 6: Ancestor Status of Multiple DNA Result Trees at mid-2010



Map originally created using software supplied by Click2Map at www.click2map.com

Map 1: Origins of West Country Trees in the Pomeroy Project

than one movement of family members from Cornwall/Devon into Dorset more than three centuries ago.

The most interesting genetic family are the trees bearing the crimson red markers (genetic family B). The tree which stands out is the one in east Devon, in the old port of Brixham. While the oldest currently documented ancestor in this tree was born in the 1780s, there are records of name bearers in the village during the previous two centuries. Most intriguing of all, Brixham is just a few miles from the ancestral heartland of the surname, the parish and castle of Berry Pomeroy just east of Totnes.

The Dorset-origin tree in genetic family B is also interesting because the enormous and well-documented tree in the USA stemming from a single emigrant in the 1630s, is believed to have come from exactly this area. While the US-emigrant tree pre-dates the Dorset tree in genetic family B by a century or more, it may be that the connection back to England will be found here.

The Clustering Process

Analysing the map is the first stage in organising a tree clustering process, which I define as a linkage process which combines genetic and documentary data with contextual evidence available to the project manager. Note that I do not see the clustering process as an attempt to be definitive, simply as a way of clarifying the project manager's thinking and helping to create fresh research priorities. There will in any case be some trees which could be placed in more than one cluster, and questions will always hang over those trees with no associated DNA result.

While the clustering process relies a lot on the project manager's general understanding of how all the data in the combined project could best fit together, certainly at the outset of phase 3 many clusters will be built around the remaining genetic families and trees with unique DNA signatures originating close by.

Table 7 below lays out the four key clusters currently defined within the Pomeroy project with the rationale behind their creation. These account for 42 trees linking roughly three-quarters of historical name bearers of British origin.

Table 8 below gives some detail about the Cornwall cluster identified in Table 7 to show how this process can work to define what is potentially a single tree using multiple types of linkage.

Cluster	Description	Linkage Rationale
Totnes (E Devon)	7 mainly small trees originating within a few miles of Berry Pomeroy castle, one of which is the Brixham tree in genetic family B, a total of 18 trees altogether.	The cluster is defined by its proximity to the castle south-east of Dartmoor and midway between Plymouth and Exeter. Genetic family B, linked via the Brixham tree, covers 9 trees and almost one in five historical name bearers.
Cornwall	11 trees originating in Cornwall, the earliest in 1575, plus 3 others through genetic linking.	Built around the 6 trees in genetic family A, which includes 3 non-Cornish trees, and other Cornish trees by proximity.
Exeter (NE Devon)	5 trees, including 3 within genetic family C.	Two trees with unique DNA signatures are linked to the genetic family by proximity.
W Dorset	Comprising 5 trees of which 4 are in genetic family E, the earliest dating from 1752.	Linked mainly through DNA results. It is possible that this cluster may prove to be a sub-set of an older Dorset tree.

Table 7: Defining Key Tree Clusters

Current Parish Origin of Tree	Year	Historical Name Bearers	Gen. Fam.	Current Rationale Within The Project
St Neot	1575	626	A	Well-established family & property owners in the 1550s; a minor branch of the noble family?
Linkinhorne	1577	1,117	u	Five DNA results, only one of which tests a line of descent free of known illegitimacies
St Gluvias	1663	89	-	Close to the important port of Falmouth; may have US descendants via several emigrants
St Neot	1689	166	u	Surely an illegitimacy within the earlier St Neot tree; several options for the father
Gorran	1717	1,079	A	DNA linkage; village 25 miles SW of St Neot
Bodmin	1747	341	A	DNA linkage; 6 miles W of St Neot
Beaminster (Dorset)	1747	139	A	Independent migration from Cornwall to Dorset, or faulty documentation?
Luxulyan	1798	89	u	Suspected illegitimacy in the Bodmin tree
Plymouth (Devon)	1808	109	A	Port and migration centre for Cornish families
Gorran	1828	114	-	Suspected illegitimacy in the earlier tree
Polruan	1828	71	u	Port lying between St Neot and Gorran
St Pancras (London)	1854	110	A	DNA linkage, documentary gap

Table 8: Structuring a Typical Tree Cluster

The above cluster, which expands genetic family A and gathers together 12 currently distinct trees around a geographical focus in the county of Cornwall, links together 4,050 historical name bearers, about 30% of the estimated historical total in the surname project as a whole. While the rationale cited for each connection varies, the overall picture the cluster creates is plausible and ready to serve both as an hypothesis and as a guide for future documentary research. Note that for some trees, DNA testing is still a work in progress.

Next Steps: Moving beyond the existing data

The 41 pins in Map 1 will be reduced to about half that number when the five existing genetic families are finally each resolved down to a single tree. Looking forward, it will be interesting at that point to add an additional layer of data onto the map, the distribution of the more than fifty manors and landholdings that the Norman-era Pomeroy noble family is recorded as holding in Devon and Somerset in the Domesday Book of 1086.[13] I anticipate that the match between the land holdings and the origins of the remaining trees will be close. The question the project has to answer is: did these trees arise in those same areas because a retainer of the family took the name by association (or likewise a villager in one of the several villages named after the family), or did they arise as younger members of the noble family slipped down the social ladder and merged into wider society, in some cases due to illegitimacy?

By the completion of the DNA strand within the dual approach project we'll have mapped a set of unique DNA signatures. What does their uniqueness signify? At one end of the spectrum of possibilities they could be a set of non-paternity events in the medieval period, perhaps including deliberate adoption of the surname along the female line as well as births out of wedlock or infidelities within marriages. At the other end of the spectrum they also fit the pattern that would be seen if unrelated retainers adopted the surname at different times and places, each one bringing their own different genetic legacy into the surname, or if name bearers adopted the surname as an alias and later dropped their original surname. Perhaps the truth, in this surname's case, is a mix of all of these scenarios. Such speculation has a role in the making of hypotheses, but the key point is that the direct evidence provided by the DNA results will always lead researchers back to a set of questions that can only finally be solved by documentary evidence.

Conclusions: 12 Summary Thoughts For Dual Approach Surname Reconstruction Project Managers

1. Documentary reconstruction back to the 1840s for surnames of English or Welsh origin is feasible because the civil registration and census data are now readily available online.
2. Documentary reconstruction is best undertaken for surnames up to a certain frequency. Using the Pomeroy project, with around 2,500 living name bearers in the UK, as a baseline I estimate that tackling surnames up to twice that frequency using the dual approach is feasible and that the method described here is therefore a viable option for all but the 1,500 most common UK-origin surnames.
3. The documentary project strand will never be completed: there will always be unlinked records and unaccounted gaps in the profiles of members of every tree. What matters more is that the documentary research defines the boundary of a dual approach project by identifying all the trees existing within the surname(s) back through time.
4. A surname reconstruction project starts and finishes with documentary research, making sense of the 'genetic families' created by the DNA results to produce a set of combined DNA & documentary data that is internally consistent and externally plausible.
5. It is methodologically better practice to separate the DNA results of testees resident outside the country of origin from those living in it; instead of mixing the data together into a single pool, use the emigrant-recipient country data as part of a separate exercise to identify the emigrant ancestors per country for each tree.
6. Though DNA testing is able to confirm the genetic inheritance of a specific ancestor within a tree, there is generally a limit how many generations back a common ancestor can be proven using DNA evidence. This often creates a gap of several centuries back to the time when surnames were founded when documentary evidence alone is available.
7. The importance of contextual data and mapping to reveal patterns and connections increases as a dual approach project progresses. The contextual data

[13] "The Domesday Book of 1086 records that the family held 57 manors in Devon, 6 houses in Exeter and two manors in Somerset. While the English royal family lost control of Normandy by 1204, the [Pomeroy] family continued to expand gaining an estate in Tregony, in Cornwall, by 1213. The earliest reference to a residence at Berry Pomeroy is in 1293 during a royal survey of the Pomeroy estates." [Source: PFA Annual Report 2009, Chris Pomery].

needs to be carefully recorded so that it can be properly challenged as new data becomes available.

8. The UK population of the Pomeroy group of surnames is estimated to have grown in size 10-12 times since 1500. I suggest this as a rule of thumb until data from other projects is available.
9. The approximately 450 present-day surname-bearing UK-resident couples collectively fit into trees associated with 25 different DNA signatures. Put another way, there is on average roughly one DNA signature per every 60 living adult males.
10. As dual approach projects develop, the number of historical name bearers associated together first into genetic families and later as unique DNA signatures will increase. Within the Pomeroy project, still only in phase 2, the three leading DNA signatures are collectively associated with trees that contain almost half of all historical name bearers.
11. As the final phase of medieval-era documentary research begins, grouping trees together into clusters concentrates the percentage of name bearers included even further. The four clusters currently hypothesised in the Pomeroy project link 42 trees and around three-quarters of all historical UK name bearers. While that level of concentration in itself does not prove that the surname is was founded by a single ancestor, it leaves the door open for that conclusion subsequently to be demonstrated.
12. All DNA projects with a parallel documentary research programme will end up in the same place: trying to make sense of the pattern of unique DNA signatures found. Does each present a non-paternity event of some kind within an existing tree, or the creation of a new tree by a specific name-taking ancestor at a particular point in time and place?

Disclosure

Chris Pomery has a commercial contract to promote Family Tree DNA in the UK. The opinions expressed in this article are entirely his own.

Acknowledgments

My thanks to Debbie Kennett, John Creer and Susan Meates for comments on a pre-publication draft of this paper.

Definitions of terms used in this article

Tree	A group of individuals, documented together and sharing a surname (or variants), which includes at least one living adult male.
DNA signature	The haplotype, or group of closely related haplotypes, which collectively define the Y-chromosome 'signature' of a particular tree or genetic family.
Genealogically-relevant timeframe	All men alive today descend from a single male ancestor. Genealogists focus on demonstrating relationships based upon common ancestry within the past millennium or so, the priamry signal of which is a shared surname.
Genetic family	A group of trees whose ancestors share the same DNA signature
Contextual evidence	Evidence, from internal and external sources and not specifically of a documentary or genetic nature, that suggests linkages between trees.
Tree cluster	A group of trees that it is hypothesised are linked together based upon a range of evidences and conjecture

Web Resources

British Telecom	www.bt.com
Family Names of the UK project	www.ahrc.ac.uk/News/Latest/Pages/familynames.aspx
Family Tree DNA	www.familytreedna.com
FreeBMD	www.freebmd.org.uk
FreeREG	www.freereg.org.uk
GenDocs (population figures)	http://homepage.ntlworld.com/hitch/gendocs/pop.html
GENUKI	www.genuki.org.uk
Guild of One-name Studies	www.one-name.org
Office of National Statistics surnames list	www.taliesin-arlein.net/names/search.php
Online Parish Clerks (Cornwall)	www.cornwall-opc.org

References

Jobling M & King T (2009), Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames, Mol Biol Evol. 26:1093-102.

King T & Jobling M (2009), What's in a name? Y chromosomes, surnames and the genetic genealogy revolution, Trends Genet. 25, 351-360.

King T et al (2006), Genetic Signatures of Coancestry within Surnames, Curr. Biol. 16, 384–388.

Pomery C (2009), The Advantages of a Dual DNA/Documentary Approach to Reconstruct the Family Trees of a Surname, Journal of Genetic Genealogy, 5(2):86-95.

Redmonds G (1997), *Surnames and Genealogy: A New Approach*, NEGHS, Boston

Sykes B & Irven C (2000), Surnames and the Y Chromosome, Am. J. Hum. Gen. 66:1417-1419.