JoGG Journal of Genetic Genealogy

Journal: www.jogg.info
Originally Published: Volume 6, Number 1 (Fall 2010)
Reference Number: 61.010

Y-DNA PROJECTS: TOWARDS IMPROVEMENT IN Y-DNA SURNAME PROJECT ADMINISTRATION

Author(s): James M. Irvine

Towards improvements in y-DNA Surname Project Administration

James M. Irvine

Abstract

This paper surveys a sample of 12 y-DNA surname projects, selected to reflect a variety of features, with the objective of identifying some possible learning points for amateur project administrators. The survey identifies a wide variety of procedures now being used in administering such projects. Many of these variations appear to reflect differing opportunities and constraints for individual projects that are determined by surname size.

The paper develops three inter-related issues directly arising from the survey. First, means for relating project size to surname size are explored. It is shown that few projects exceed a "penetration" ratio of more than 0.1% of y-DNA tests per head of population, and that this ratio may be an inverse function of surname size. Measures are also developed to relate old world/new world ratios of surname populations and participants' places of residence; from these a crude measure of any geographical bias in individual projects is developed. Second, the survey identifies a diversity of the "rules of thumb" presently used for determining genetic "closeness," and a case is made for moving on from genetic distance criteria that give equal weight to all markers to a criterion that takes account of differing mutation rates, such as some TiP parameter. Third, the difficulties in identifying and handling the sensitive subject of Non Paternal Events (NPEs) are addressed. A case made for differentiating between introgressive- and egressive-NPEs, and it is shown that most projects probably underestimate this phenomenon. A brief summary of the Irwin surname project is appended.

Introduction

For the past decade geneticists and genealogists have been collaborating in DNA surname projects to exploit the shared characteristic of both y-DNA and surnames that they generally pass unchanged from father to son. The two disciplines have struggled to grapple with the significance of haplogroups, haplotypes, random mutations, mutation rates, clusters, NPEs, variations in surname spellings, and the many vagaries of genealogical records and indexes, both private and public. The skills of statisticians, data managers and webmasters are also needed. And the rapidly evolving genre has to compete with parallel interests in SNP, mt- and autosomal DNA tests, and in deep ancestry studies.[1]

Academics have brought some light, if limited impact. Even before the advent of y-DNA tests it was recognised that the traditional derivations of surnames from place names, personal names, trade names and nicknames were simplistic:[2] the origins of non-hereditary names, alias names and anglicised surnames need more attention than they have been given hitherto, and it is becoming increasingly clear that DNA evidence should be taken into account before attempting to classify the type and meaning of each surname. Another topical issue is whether individual surnames have single, plural or multiple origins. Intuitively, at least place- and trade-surnames should have multiple origins. But DNA surname studies by Sykes & Irven[3] and Pomery[4] have suggested that single-origin surnames may be more

Address for correspondence: James M. Irvine, jamesmirvine@hotmail.co.uk

Received: June 22, 2010; Accepted: Nov. 14, 2010; Published: Dec. 17, 2010

[1] For the purposes of this paper I consider "deep ancestry" to include all applications of DNA tests outside the timeframe of conventional genealogy, i.e. before surnames became hereditary, generally at most about a thousand years ago.

[2] George Redmonds, *Surnames and Genealogy, A New Approach* 1997, 1-17.

[3] Bryan Sykes & Catherine Irven, *Surnames and the Y Chromosome* 2000: www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1288207.

[4] Chris Pomery, *Family History in the Genes* 2007, 210-1.

common than had hitherto been suspected. King & Jobling[5] have suggested that single-origin surnames are often associated with the less common surnames, while McEvoy & Bradley[6] have shown many common Irish names have a single major ancestor, implying geographical origin may also be relevant. Meanwhile Plant[7] has introduced some useful concepts to help us understand and handle the phenomena and implications of NPEs.

In parallel with such academic studies, and potentially overshadowing them, commercial developments have led to a dramatic growth in the number of y-STR DNA test participants and surname projects. The results of well over 200,000 such tests are now available on the web, listed in the public databases of testing companies[8] and by about 6,000 surname projects hosted by FTDNA, WorldFamilies and various private websites. The size distribution of these surname projects is revealing, as shown in Table 1.[9]

Indeed many of the smaller surname projects have less than 10 participants, while fewer than 100 such projects have more than 200 participants. But steady growth in the numbers of participants and projects continues.

The interpretation of this vast amount of data has been in the hands of volunteer administrators of individual surname projects[10] who can draw on advice and suggestions of the testing companies and exchange views at conferences and seminars organised by the testing companies, the Guild of One Name Studies and others, a variety of journals such as JoGG, and web discussion groups such as ISOGG. Such is the rate of development of the testing facilities, the understanding (or lack of) the underlying science, and the absence of constraints such as the peer reviews of academia, that surname projects have developed with little common terminology or strategic co-ordination. This is not the time or place, nor I the person, to remedy these deficiencies, and the purpose of this paper is more modest: I am simply attempting to compare some surname projects and to address a few of the issues arising, with the objective of improving the awareness of other project administrators, and so hopefully helping us to grapple with some of the many emerging challenges.

In any attempt to analyse and review individual surname projects, several important caveats have to be noted:

- While companies such as FTDNA, Ancestry.com and others undertake the y-DNA testing, the administrators of DNA surname projects in which intra-surname results are grouped and analysed are all volunteers, with varying individual constraints of knowledge, skills and time.
- Surnames have very different characteristics, so deviations from “norms” must be anticipated and respected. Each project has its own goals and constraints (e.g. population size, availability of funds, access to genealogical data, administrator limitations), and comparisons can be odious.
- A comprehensive analysis of all the surname projects is impractical, while a small sample, such as addressed in this paper, cannot be considered representative.
- Many features of individual projects are subtle, and differences in presentation and content, though important (particularly in attracting new participants), are not amenable to objective comparisons. Much work “behind the scenes” remains unpublished, and nuances cannot be developed in a paper such as this.
- Although analysis of published DNA surname project data with full academic rigour is neither possible nor appropriate, care is necessary to ensure all comparisons are on a “like-with-like” basis as far as possible.
- Testing companies and project administrators adopt a bewildering variety of terms to describe genetic, genealogical and statistical terms.[11] Here I have preferred to adopt non-technical terms that are as self-explanatory as possible for the newbie.[12]

[5] Turi King & Mark Jobling, ‘Founders, drift & infidelity: the relationship between Y chromosome diversity and patrilineal surnames’ in *Molecular Biology and Evolution* 2009, v.26, 1093-1102, reprinted as http://mbe.oxfordjournals.org/cgi/content/abstract/msp022, 1-38.

[6] Brian McEvoy & Dan Bradley, *Y-chromosomes and the extent of patrilineal ancestry in Irish surnames* 2006: www.ncbi.nlm.nih.gov/pubmed/16408222.

[7] John S Plant, *Surname studies with genetics* Guild of One Name Studies 2009: http://cogprints.org/6595/.

[8] For details of y-DNA tests offered by the five main testing companies see www.isogg.org/ydnachart.htm. FTDNA have completed over 180,000 y-DNA tests and Sorenson over 35,000. Equivalent data for other companies is not available, but I see no reason to doubt FTDNA’s claim that its database is larger than those of all its competitors combined.

[9] From http://www.worldfamilies.net/surnames as of 1st April 2010. I deliberately limit the percentages in this and many of the following tables to two significant figures, which is often all that the data justifies, or that the point being made requires.

[10] Project websites use the terms group, study & project, and administrator & co-ordinator interchangeably. Here I use the most popular combination, project administrator.

[11] Charles Kerchner captures many of these in his dictionary: http://www.kerchner.com/books/ggdictionary.htm.

[12] For example I prefer genetic signature, or possibly genetic fingerprint or profile, to haplotype, marker count, motif, or repeat, though I recognise these terms are not necessarily synonymous, and the preferences of others may differ.

No. of participants	% of surname projects	
0- 99	95%	of which 89% have < 50 participants
100-199	3.7%	
200-299	0.7%	
300-399	0.4%	
400-599	0.12%	Baker, Davis, Graves, Harris, Phillips, Rose
600-799	0.08%	Johnson, Williams
800-999	0.03%	Clan Fraser, Clan Donnachaidh

Table 1

- The use of percentages reduces the need for frequent updating of data as additional test results emerge, and facilitates inter-project comparisons. But to retain accuracy it is sometimes necessary to use different denominators within the available data sets.

- Surname project data can age quickly, and some project websites are only updated infrequently; in this paper I have used data as available on 1 April 2010, sometimes supplemented with subsequent personal advice from project administrators.

- The "science" of y-DNA is still evolving rapidly, and dogmatic conclusions, even when both possible and desirable, may be premature.[13]

Bearing these points in mind, for this comparative study I have selected, somewhat arbitrarily, twelve y-DNA surname projects with a variety of characteristics and which hopefully include the work of some of the more innovative administrators. The selected projects are, in ascending order of surname size:

Creer, Pomeroy, Plant, Cruwys, Dalton, Blair, Irwin, Phillips, Wright, Walker, Taylor and Williams.[14]

My inclusion of these projects, and exclusion of others, in no way implies judgements of the relative "quality" of any particular project or administrator – my objective, as already explained, is simply to seek some possible future improvements in understanding, interpretation and context for all project administrators.

1. Size, growth, population and penetration

Intuitively, the number of participants in a surname project is one simple measure of its success. This is particularly so with DNA test results, as mutations are random and test results need to be considered on a statistical basis. Even if a pedigree is fully traced, a single DNA test is not necessarily representative of that pedigree, for although the genetic distance between full brother and father/son DNA signatures is usually 0, it can, very occasionally, be 2 or 3. So for a given surname, from a reliability point of view, clearly 200 test results are better than 20. On the other hand large surname projects incur considerable cost (even if funding is available) and administrator time and workload. Indeed at the smaller end of the size spectrum, growth in DNA tests can detract from the prime objective of using DNA as a tool to help genealogy.[15]

2.1 Advertised size vs. number of y-DNA test results analysed

One measure of project size, the number test kits issued by FTDNA to participants in a project, is readily available on the FTDNA and WorldFamilies websites. These "advertised" project sizes, as of 1 April 2010, are listed for our selected projects in Appendix A, line 4.
From these sizes, and the date each project was founded, I have calculated the average growth rate of each project (line 6). These rates range from a new participant every three days to one every two months. However such rates imply linear growth, which may be misleading as some projects grow relatively fast initially and their rate of

[13] I must also stress that (1) this paper is not intended to constrain in any way the on-going development of the rapidly evolving genre of DNA surname projects, nor to compromise the confidentiality of individual participants' data, and (2) it is inevitable that the ideas advanced are a compromise between academic rigour and something simple enough to be readily understood and implemented by most project administrators.

[14] Creer: http://www.creer.co.uk/.
Pomeroy: http://www.one-name.org/profiles/pomeroy.html.
Plant: http://www.plant-fhg.org.uk/dna.html.
Cruwys: http://www.familytreedna.com/public/CruwysDNA/default.aspx.
Dalton: http://www.daltongensoc.com/dnaproject/index.html.
Blair: http://blairdna.com/.
Irwin: www.clanirwin.org > DNA Study.
Phillips: http://www.phillipsdnaproject.com/.
Wright: http://www.wright-dna.org/.
Walker:http://www.familytreedna.com/public/Walker%20DNA%20Project%20mtDNA%20Results/default.aspx.
Taylor:http://www.familytreedna.com/public/taylorfamilygenes/default.aspx.
Williams: http://williams.genealogy.fm/dna_project.php.

[15] Chris Pomery makes this point in 'The Advantages of a Dual DNA/Documentary Approach to Reconstruct the Family Trees of a Surname' in *Journal of Genetic Genealogy* 2009 Fall, 91, 92. The Creer project was even more selective, only including participants already identified from genealogical research.

growth then slows. Rate of growth is a very crude indicator of administrator workload, but this workload also varies considerably from participant to participant, and from project to project, depending on how ambitious their goals are. Larger projects can necessitate sharing and delegating of the workload amongst two or more administrators. Some projects (e.g. Pomeroy, Dalton and Taylor) have established separate administrators and webpages for individual clusters. Other projects may be just one of as many as fourteen managed by a single administrator.

But the advertised project size usually differs from the number of test results that administrators use in their own analyses (Appendix A, line 8). Advertised sizes may include unused test kits held by administrators, and test kits that have been sent to participants but not yet been returned, and even some mt-DNA and autosomal DNA test results as well. There is also usually a lead time, sometimes of several months, before administrators publish and analyse their test results. On the other hand the advertised sizes exclude test results that many administrators include in their projects which have come from testing companies other than FTDNA. These differences thus fluctuate over time, and Appendix A, line 9 suggests that the number of y-DNA test results analysed by a particular project may be anything from about 15% below FTDNA's advertised size (Cruwys) to 60% above (Pomeroy)!

2.2 Population and penetration

When assessing the progress and representativeness of a surname project the number of DNA tests completed is less relevant than the ratio of DNA tests per head of the population of that surname. I use the term "penetration" to describe the ratio of the number of y-DNA tests for a given surname to the world population thereof. The importance of penetration is illustrated by comparing the Pomeroy and Williams projects: although in terms of test results the Williams project is six times the size of the Pomeroy project, when the world population of these two surnames is taken into account the Pomeroy project has twenty times the penetration of the Williams project.

Penetration has been considered qualitatively before, and even numerically,[16] but hitherto attempts at quantification have been frustrated by the lack of surname population data on a global basis. However the calculation of approximate populations of individual surname spellings in each country, and hence the world has been enabled by the University College of London's publicprofiler website.[17] Although this database has some important limitations (discussed in Appendix B), it enables surname project administrators to calculate indicative world population figures for different spellings of their surname for the first time.

The calculation is explained in Appendix B, and the resulting approximate populations for the selected surnames are listed in Appendix A, line 11.[18] The following categorisation of surname sizes may be appropriate:

Given the world population for each surname, calculation of the penetration of each project is straightforward:

Penetration % for each surname =

$$\frac{\text{No.of completed yDNA tests for the surname x 100}}{\text{World population of the surname today}}$$

The resulting penetrations for the selected projects are shown in Appendix A, line 13. The mode of these penetrations is 0.06%: in other words, the number of y-DNA tests undertaken within most surname projects is still less than 0.1% of the world population of their surname. Although penetration ratios are steadily improving, it is important that administrators recognise that this is clearly still a poor sample rate in terms of conventional surveys in other fields,[19] and, more relevant, a very poor rate if one of the project goals is to find living cousins, or to identify the DNA signature of all the branches of a surname.[20]

Individual project penetration rates range between 0.02% and 1.5%. Clearly such rates will have been influenced by factors such as how long the project has been running, the availability of funds to individual projects and hence the extent of pro-active recruiting of participants (whether by administrators seeking under-represented

[16] e.g. maps of 19th and early 20th century UK and USA census data showing surname distribution by area. Plant 2010 and others have used the UK Office of National Statistics data (http://www.taliesin-arlein.net/names/search.php), but this only covers England and Wales. Pomery has used private estimates. The Dynastree database is considered in Appendix B above.

[17] http://worldnames.publicprofiler.org/.

[18] The resulting population figures are probably only accurate at best to three significant figures. But I have not rounded them thus so that the subsequent calculation processes remain clear and avoid consequential errors.

[19] In fact the situation is not quite as bad as these figures imply, for only males can take y-DNA tests, and although FTDNA do not restrict tests to adults, effectively only the adult male population is eligible. In other words in practice "penetration" cannot exceed about 40% of the total population of a surname. Penetration rates may thus be divided by 0.4 if a more realistic "feel" for potential penetration is required. But even on this basis few projects have to date have achieved more than 0.2% of their potential penetration. NB I have omitted this factor of 0.4 from my definition of penetration to keep it simple.

[20] By a different measure of penetration the Creer project has proactively tested representatives of 75% of the pre-identified branches of the surname.

World population of surname	< 5,000 Very small	5,000 - 50,000 Small	50,000 - 500,000 Medium	> 500,000 Large
example projects	Creer	Pomeroy	Plant[21] Cruwys[21] Dalton Blair Irwin	Phillips Wright Walker Taylor Williams

[21] The Plant and Cruwys projects include of several like-sounding but apparently unrelated surnames, such as Plante & Plants, and Cruise & Crews.

Table 2

branches, or participants seeking close relatives), "pre-entry" requirements (e.g. Creer, Blair), and competition from such projects run by rival testing companies (e.g. Pomeroy, Taylor).[21] Such biases are not explored in this paper. However, as might be expected, there is apparently some correlation between penetration and population size for a given surname: of the selected projects, the smallest surname has the largest penetration (where the rewards of combining DNA results with conventional genealogy have attracted participants, and/or funding has enabled pro-active testing of selected participants?), while the largest surnames have the smallest penetrations (participants deterred by a suspicion that DNA is unlikely to add so much to the genealogies of trade and personal surnames?).[22]

It would be inappropriate to assume these findings apply generally, but they do nevertheless underline the importance of the concept of penetration. And of course while penetration is a measure of the "quality" of a particular project, a high penetration is only one of several factors that contribute to a "good" project, and a low penetration does not necessarily infer a "bad" project.

2.3 Surname spellings and geographic distribution

In calculating the population of a surname and the penetration achieved by the associated project, some assumption has to be made on which spelling variants[23] should be included when calculating penetration. All the spellings used by the participants should be included, plus other spellings that would be acceptable to the project. The number of spelling variants I have used to calculate surname populations are indicated in Appendix A, line 10. The wide range of the number of variants so used, from 1 to 35, is less dramatic than it may appear, as typically the less common variants add little to the total populations. Indications of only a single variant reflect my understanding that the relevant project only accepts participants with that spelling.

One important benefit of the matrix developed in Appendix B is the opportunity it gives project administrators to see how the spread of different surname spellings is now distributed around the world. The matrix thus gives the best available overview of the global distribution of the places of residence of potential project participants, i.e. it is a valuable tool for understanding the evolution and diaspora of surname spelling variants, and an indication of how well the associated project is addressing this dimension and how it may develop in the future.

2.4 Geographical distribution of surnames today, & Old world/New world ratios (Appendix A, lines 14-16)

The matrix in Appendix B also enables the calculation of ratios of the populations of individual surnames resident in different countries, for example UK to USA. A more general ratio, which I call "population drift,"[24] is the proportion of the population of the surname resident in the immigrant-receiving countries of the New world[25] to the total population of the surname, i.e.

$$\text{Population drift } \% = \frac{\text{New world population of surname today x 100}}{\text{Total world population of the surname today}}$$

A manifestation of population drift is the population of the Old world today being appreciably less than that of the New world, even though much of these populations share common ancestry. Population drift ratios are

[21] My survey is restricted to surname projects listed by WorldFamilies, i.e. generally dominated by FTDNA data.

[22] Adrian Williams (pers. comm. June 2010) has shown that the low penetration of the Williams surname in Confederacy States in USA may be due to emancipated slaves taking the surname of their former owners, but few of the descendants of these slaves participating in the Williams surname project.

[23] Derek Palgrave has differentiated between surname variants (genuine spelling variations) and deviants (transcription errors) ('Many surname variants are really misspelt deviants' *Journal of One-Name Studies* Jan-March 2004, 6-9). Alas in the present context we need to address both.

[24] I have adapted this term from the geneticist's term genetic drift, as the latter relates to random changes in genetic signatures from generation to generation, whereas my term "population drift" includes genetic drift as well as natural selection, migration rates, "bottlenecks", post-migration birth and mortality rates, and social and economic factors.

[25] This split into Old world/New world terminology, already adopted by the Dalton project, is thus intended as a crude measure of migration element within a surname, i.e. its diaspora. The vast bulk of the New world participants are likely to reside in USA compared with Canada and Australasia, but it would be misleading to assume that USA residence ratios alone are an optimal indicator of migration from the Old world.

available for seven of our selected surnames, and have a mode of about 80%. The ratios range from 68% to 86%: seemingly relatively few Creers and Plants migrated, and/or those that did migrate propagated at a slower rate than other immigrants. In fact many genealogists suspect that migrant families generally procreated faster than their homeland relatives, but I am not aware of formal research having addressed this feature.

Population drift, as defined above, has nothing to do with DNA testing. But project administrators can derive similar ratios for their participants by their countries of residence. For this I use the term "project drift" to identify the ratio of participants in the surname project resident in "New world" countries to the total number of participants in the project, i.e.

$$\text{Project drift \%} = \frac{\text{No. of participants resident in New world x 100}}{\text{Total no. of participants in project}}$$

The project drift ratios of the seven surnames range from 64% to 94%, but the mode seems to be higher than that of population drift, about 85%.[26] In other words, most surname projects have a predominance of participants residing in the New World. However the geographical bias of a particular project is not how its project drift compares with that of other projects, but how it compares with the population drift for the same surname. Thus while the Dalton project has a project drift of 82%, this is not dissimilar from its population drift of 81%. A crude measure of the geographical bias of a surname project may thus be assessed by the ratio of its project drift to its population drift, i.e.

$$\text{Project bias} = \frac{\text{Project drift}}{\text{Population drift}}$$

In theory this bias ratio should be close to 1.0, as it is for the Cruwys and Dalton projects. But for the Plant project the bias is 0.94, indicating this project includes a lower proportion of participants residing in the New world than might be expected, while the Irwin, Blair and Phillips projects have biases of between 1.06 and 1.16, indicating poorer participation in these projects by residents in the Old world than may be desired. Project bias is thus a measure of an issue, real or just perceived, that vexes many project administrators.[27]

Some understanding of the concepts of what I term population drift, project drift and project bias is important to project administrators. The primary objective of some surname DNA projects is to use DNA to help merge participants' pedigrees. In extremis the goal of a single-surname DNA project is to create a single family tree that includes all the participants. As the main challenge in achieving that goal lies with the handling of early, "Old world" pedigrees, there can be an understandable lack of enthusiasm by some to attract "New world" participants. But excessive emphasis on Old world participants risks overlooking the possibility of some New world participants representing lines that have become extinct in the Old world since migration occurred. It also risks alienating New world participants who may be willing to donate funds to subsidise DNA tests for potential Old world participants who might otherwise be reluctant to undergo such tests because they are apprehensive about confidentiality issues, or who do not perceive the cost/benefit of a test justifying their involvement.

Conversely some projects, especially those of larger, multiple surnames, seem more interested in attracting participants resident in New world countries, and struggle to make their projects appeal to potential participants resident in the Old world whose DNA signatures and early pedigrees could be a major contribution their project.

3. Paper trail data (family trees, pedigrees, lineages, patriarchs)

The importance of complementing DNA research with conventional genealogical research, what some geneticists term "paper trail data", is rightly stressed by Pomery.[28] In practice the extent to which individual projects are using DNA and paper trail data as complementary sources varies widely, reflecting differing goals and objectives. Project goals generally include the use of DNA to:

1. Complete genealogical pedigrees for all with the surname, and identify all the surname founders, in extremis the single-origin family: only practical for small surnames, typically UK based.

2. Break down "brick walls", and find cousins: typically for very large, multi-origin surnames, US based.
3. As 2. above, plus to identify the DNA signature of pedigrees of known ancient families and

[26] The much lower population drift and project bias ratios for the Pomeroy project are not directly comparable because this has a deliberate policy of focussing on Old world participants.

[27] The reasons why some projects struggle to attract participation in the Old world is beyond the scope of this paper but see, for example, exchanges on 'DNA Testing Company for British customers' and 'Phillips' DNA Project' at http://archiver.rootsweb.ancestry.com/th/index/GENEALOGY-DNA/2010-09 on September 19-23.

[28] Pomery 2009, 86-95.

trace geographical origins of migrant families: typically for medium and large surnames.

4. Achieve more specific goals, such as to establish the modal DNA signature and likely geographical origin of each branch of the name, or determine whether the surname is single-origin or plural-origin.

To meet their goals project administrators adopt various approaches to handling paper trail data. All FTDNA participants can post their GEDCOM trees for viewing by other participants in their project. In addition most administrators seek to list the male pedigrees of every participant, although some (e.g. Plant and Irwin) only publish details of their earliest known paternal ancestor. Details of the earliest known ancestors in the Creer, Pomeroy and Dalton projects are only available to non-participants via the ysearch website. Some administrators use their knowledge of early genealogies to target potential new participants.

There is some recognition that the quality of data on earliest known ancestors can be questionable:[29] the Blair project differentiates between "ancestors" and "oldest ancestors" and the Irwin project between the "earliest confirmed paternal ancestor" and the "traditional geographical origin of this individual's ancestors", while the Williams project (and no doubt some others) tries to check the authenticity of each participant's ancestry.

Interesting comparisons can be made of the geographical origins of the earliest known ancestor (Appendix A, lines 17-24), and the dates they lived (lines 25-28). Unless prepared by the project administrator such statistics are laborious to compile, but those available show that the proportions of participants able to trace their ancestry to the Old world range widely, from 5% (Williams) to 89% (Pomeroy). For each of the seven projects analysed by date of earliest ancestor, about half of their participants able to trace their ancestry back to before 1800, except for the Williams project where only about a quarter can.

3.1 Average male generation interval

An incidental feature of this important attention to paternal pedigrees is clarification of the average male generational interval. Geneticists traditionally advise 25 years per generation,[30] but this may include the shorter intervals associated with females and/or studies more relevant to deep ancestry. For male generational intervals during the past millennium Devine quoted historical studies of 31-38 years, 35, 32 and 34 years.[31] King & Jobling adopted 35 years.[32] The earliest pedigree in the Cruwys project shows average intervals of 30-35 years. The Williams project finds 28-33 years.[33] My own research shows nine Irwin pedigrees dating from between 1323 and 1660, with a wide range of socio-economic backgrounds, have intervals of between 31 and 38 years.

All the above discussion relates to the context of y-DNA testing, but we can now move on to explore how our selected projects handle their test results and the issues arising.

4. Publication of y-DNA test results

Because of early undertakings given on confidentiality, individual test results in the Creer, Pomeroy and Dalton projects can only be viewed by non-members on the ysearch web pages. Participants' test results in the other selected projects are published on the project web sites.

5. Resolution (Appendix A, lines 29-33)

All the selected projects recognise the higher resolution the better. Attitudes to the older 12-marker tests vary, from disparaging to them being at least "a foot in the door."

Resolution statistics are available for all 12 of the selected projects. All but the Taylor and Plant projects now have at least 60% of their participants with 37 markers or more. All the participants in the Cruwys project have 37 or 67 markers, in part because of its later start-up date.

6. Definitions of close matches, clusters, genetic families, singletons and modal DNA signatures

Perhaps the most important role of each project administrator is to sort the test results of the projects' participants into closely matching clusters, also known as groups (a term also used by FTDNA to denote projects), subgroups (FTDNA), family groups (Phillips), lineages (WorldFamilies, a term also sometimes used by others to denote pedigrees), genetic families (Pomeroy and Irwin) or branches. Those participants whose results don't "match" with any other of the surname are termed

[29] This is particularly so if the paternal pedigree has been complied solely from opportunistic searching of IGI data.

[30] Bruce Walsh even has 15-25 years (http://nitro.biosci.arizona.edu/ftdna/models.html#Time).

[31] Donn Devine 'How Long is a generation?' *Ancestry Magazine* 2005: http://www.ancestry.com/learn/library/article.aspx?article=11152.

[32] King & Jobling 2009, 1095, http://mbe.oxfordjournals.org/cgi/content/full/26/5/1093.

[33] Debbie Kennett, pers. comm., March 2010; Adrian Williams, pers. comm., May 2010.

"Unassigned" by FTDNA but "singletons" by many administrators. Here I use the terms cluster and singleton.[34]

Curiously there is not only little consensus on this nomenclature, but also on the processes involved. Each administrator has determine, consciously or unconsciously, how he is going to:

- ascertain whether or not any two participants should be considered as a "close match";
- define and identify each cluster;
- assign each participant to the relevant cluster or to singleton status;
- decide how to sequence participants within each cluster; and, for many
- determine the modal signature of each cluster.

Defining what does or does not constitute a "close" or "near" match between two participants of the same surname is essentially a subjective judgement based on probabilistic genetic data and often genealogical material as well. So it is neither surprising nor unhealthy that a rigid definition is elusive. Nevertheless administrators do need some "rule of thumb", consistent at least within their own project. For this purpose all the administrators of the selected projects appear to draw on the probabilistic guidance offered by FTDNA on genetic distance.[35] Most apparently seek interpretation at a 50% probability level.[36] But administrators adopt a surprising variety of the "rules of thumb" for their criteria to determine whether or not two participants are a "close match", and to assign a participant to a cluster:

- The Walker project follows current FTDNA advice, which can be interpreted to infer that two participants with the same surname have a match or "near match" if they have genetic distances up to 1 at 12 markers, 2 at 25 markers, 4 at 37 markers, and 7 at 67 markers. I refer to this as a "1, 2, 4, 7" rule of thumb.[37]
- The Creer project and the WorldFamilies web tutorial use a genetic distance up to 2 at 25 markers.
- The Cruwys project uses a "1, 2, 4" rule of thumb, augmented by paper trail data and triangulation.[38]
- The Dalton project uses a "1, 3, 5" rule of thumb.[39]
- The Taylor project uses a "0, 2, 3, 5" rule of thumb.[40]
- The Blair project uses a "0, 2, 4, 6" rule of thumb.[41]
- The Williams project uses a "1, 2, 4, 6" rule of thumb.[42]
- The Irwin project uses the individual participant's 80% probability from FTDNA's 24-generation TiP tool. The justification and benefits of this rule-of-thumb are developed and discussed in Appendix C.

It is thus clear that any comparison of statistics on clusters and singletons for different projects has an element of "apples and pears", even though in practice it would be unlikely that the adoption of some single "rule of thumb" would radically change the overall picture.

The next stage, generally accepted, is recognition that if two (three in the Williams project) participants are a close match then they form a cluster or a genetic family, which Pomery defines as a "shorthand phrase to define men that have an identical, or near identical, *haplotype* or *DNA signature* and who share a common surname. The 'genetic family' is what you create when you aggregate similar Y-chromosome DNA results together." [43] In the Irwin project I apply several caveats to this definition:

- participants related closer than second cousins only count as one participant;[44]
- a genetic family need have only one participant if his DNA test is accompanied by a paper trail or other evidence tracing his paternal ancestry back to the 16th century or earlier;
- a genetic family may include participants with different surnames if there is clear evidence of an e-NPE descent (see section 7 below).

Singletons are the remaining participants not assigned to a cluster or genetic family. They may include participants whose test resolution is too low to form a confident interpretation, and those who at some time in

[34] I am also attracted by regarding "cluster" as a generic term that covers the hierarchical concept of haplogroup, genetic family, genetic branch and haplotype.

[35] FTDNA use a hybrid definition of genetic distance: the stepwise mutation model for all alleles except DYS464 and YCA which use the infinite allele model. In the stepwise model each mutation is allowed to change the allele value by exactly one, so a difference of two means that two mutations occurred and a difference of three means that three mutations occurred. In the infinite allele model the entire difference between allele values, no matter how large, is counted as a single mutation.

[36] Presumably including "Almost certainly" and "Probably", but excluding "Possibly" and "Very unlikely"; Taylor uses 80%.

[37] See http://www.familytreedna.com/genetic-distance-markers.aspx?testtype=[12]/[25]/[37]/[67]. Perversely FTDNA uses a "1, 2, 3, 7" rule of thumb for their "Relevant matches" (https://www.familytreedna.com/privacy-policy.aspx).

[38] Debbie Kennett, pers. comm., 9 May 2010. Triangulation is discussed in section 9 above.

[39] Michael Dalton, pers. comm..

[40] http://freepages.misc.rootsweb.ancestry.com/~taylorydna/groups.shtml > What is a group?

[41] Derived from average 50% probabilities in the 12-generation TIP tool (http://blairdna.com/dna103.html).

[42] Adrian Williams, pers. comm..

[43] These terms were introduced by Pomery 2007, 224; 92, and King & Jobling 2009, 11.

[44] This refinement follows the principle set out in King & Jobling 2009, 31.

the future may be joined by another closely matching participant with whom they will form a cluster that is either:

- a genetically distant branch of a single-origin surname project; or
- a genetically different branch of a plural or multi-origin surname; or
- an i-NPE from a quite different surname (see section 7 below).

6.1 Modal DNA signatures

A subtle but radical analytical tool is recognition of the supposition that within a cluster all the participants share descent from a common ancestor. Without the use of triangulation (see section 9.7 below) it is impossible to determine the DNA signature (or name or dates) of this ancestor, but the modal DNA signature of each cluster can be readily determined.[45] So in practice the modal signature can serve as the basis for "close match" comparisons to determine whether or not any participant qualifies for membership of the cluster. It follows that there is no need to establish a matrix of probabilities of relationships between all participants in each cluster - a monumental task with large clusters.

In practice the modal signature of a cluster may be considered to be the signature of the participant(s) sharing (or closest to[46]) this modal signature. When the cluster has only two participants, or a modal marker value is unclear (e.g. three participants with one or more markers that differ for all three participants), the modal signature may be assumed to be that of the participant with the earliest confirmed paternal ancestor until a new participant is assigned to the cluster. And of course the modal signature may change as the cluster grows.

The process for assigning new participants to a cluster is thus to compare the participant's signature with the modal (or singleton) signature of his closest match to see if they qualify for cluster membership, if they create a new cluster with an existing singleton, or if, for the time being, they have to be assigned singleton status.

Modal DNA signatures are used in the Plant, Dalton, Irwin, Phillips, Wright, Taylor and Williams projects (where they are called modal haplotypes) and Blair project (ancestral haplotypes).

7. NPEs (non paternal event, false paternal event, false paternity, misattributed paternity, non-patrilineal transmission, male introgression, ancestral introgression)

There are many reasons that a male's surname may differ from that of his biological father, including:

- illegitimacy, both in-wedlock (including covert infidelity) and out-of-wedlock, when a young boy was given the name of his mother or her husband (all periods, countries, and social classes);
- formal adoptions/name changes (post 19th century, and earlier, in Scotland at least, by a husband or widower so that he could inherit land from his wife or father-in-law);
- unrecorded adoptions/name changes, e,g. when:
  - an orphan was given the name of his guardian;
  - a young boy was given the surname of his widowed mother or his stepfather (all periods, especially pre 19th century, and in Scotland where wives retained their maiden names);
  - a migration, naturalisation or administrative change led to the anglicising of a surname;
- changes in surname before these became strictly hereditary (typically 12th-18th centuries), e.g. when:
  - a boy took the patronymic of his father;
  - a boy took his mother's surname when she had higher status than his father (e.g. Oliver Cromwell);
  - a tenant, servant, apprentice, slave (USA - see endnote 24 below), or clan member (Scotland) took the name of his landlord, master or chief;
  - a man became known by his alias name (i.e. "aka"), such as his trade- or nick-name, or the name of the place from which he had migrated, or in which he now lived.

It can be seen that while some of these contingencies include ancestry through a maternal line, it is quite inappropriate to assume or infer that most NPEs are associated with illegitimacy.[47]

Plant has shown that, in the context of any individual project, two modes of NPE need to be considered:[48]

(a) introgressions ("i-NPEs") into the project surname from some other, earlier surname, perhaps even a non-hereditary name. Such

[45] The modal DNA signature <u>may</u> be the signature of the common ancestor, but not necessarily so. For example, the "founder effect" can introduce a "bias" in the project population if a small number of migrants established a new population that procreated faster than the original population left behind, thereby creating a larger but less genetically diverse population.

[46] In theory use of the signature of the participant closest to the mode could be avoided by asking FTDNA to fabricate a "theoretical" cluster modal signature (e.g. ysearch C7BED). But in practice I have not yet found any need for this.

[47] Technically speaking it can be argued that NPEs should not include patronymics, but on the other hand could include clerical errors in parish registers , and even mistakes in genealogical research.

[48] Plant 2009, 9.

participants will bear the project surname, but bear the DNA signature of some other surname.

(b) egressions ("e-NPEs") from the project surname to some other, later surname. Such participants will not bear the project surname, but will bear the DNA signature of one of the clusters of the project surname, indicating the participant's paternal ancestry before the event was that of the project. In practice e-NPE participants will probably have first joined another project which will have found them to be an i-NPE in that project. Caution must be exercised to ensure random matches are not included.

It follows that potentially one project's e-NPE will be another project's i-NPE, and vice-versa.

i-NPEs (i.e. participants bearing the project surname, but with DNA not matching any other cluster) may be singletons, i.e. they do not (yet) have any close matches, or they may form one or more clearly matching clusters within the project. The surname of each singleton or cluster before the "event" will fall into one of several categories:

(i) A surname known or suspected because the "event" was fairly recent. But in practice even if such individuals have had a y-DNA test, they are unlikely to register with the project of their "new" surname.

(ii) A surname identified by opportunistic searching in the FTDNA personal pages or ysearch pages as being a close match, and where genealogical research has suggested an "event" occurred when the two families are known to have lived as neighbours, i.e. in the same district at the same time.

(iii) A surname similarly identified, but with no apparent connection with the current surname project. Here the date of the "event" will also be unknown, even very approximately.

(iv) A surname that does not match the DNA signature of any other surname cluster. These are most difficult to interpret. They may be a singleton representing some "other" surname of which no one else has yet undergone a y-DNA test or, in extremis, a surname that is now almost extinct. In plural or multi-origin surname projects these i-NPEs are not readily distinguishable from separate branches of the name, and in practice will be treated as just another branch or lineage.[49] In single-name projects these i-NPEs may be associated with early paternal ancestry that is co-located with clusters of the same surname. Pomery argues these may reflect a single source surname with some early NPE, perhaps through a maternal line before the name became strictly hereditary, or one of the other examples above.

Most – but not all – of the selected projects recognise i-NPEs, though some administrators are apprehensive of discussing them in case participants are embarrassed. In the Irwin project 8% of the participants have been identified as i-NPEs on the basis of the criteria above and are grouped in seven clusters, all with modal signatures similar to families that originally resided in the same part of Scotland as the common ancestor (i.e. (ii) above). This feature suggests that these "events" all occurred before migration from the Borders of Scotland, i.e. probably between the 13th and 16th centuries. It also has four other clusters, each with distinctive DNA signatures, that closely reflect the tradition that different geographical branches of the name elsewhere within Scotland share a common ancestor (i.e. (iv) above).[50] Explaining that many NPE's are not illegitimacies has minimised potential embarrassment.

By definition e-NPEs (i.e. participants not bearing the project surname, but with matching DNA and having some very clear genealogical or geographical connection with the project) cannot be singletons, and their signature will be a close match with one of the project's clusters (i.e. (i) or (ii) above). Some of the selected projects seem to be unaware of them or even reject them, and apparently only the Irwin, Phillips, Wright, Walker, Taylor and Williams projects include them. But they are more tricky to handle. In the Phillips project e-NPEs constitute 3% of all participants, but they are listed separately rather than within the relevant cluster(s). In the Irwin project 15 e-NPEs are included,[51] even though they constitute 8% of the total number of participants and thus distort the project's penetration and bias ratios. The backgrounds to these e-NPEs are the reciprocals of the i-NPE categories (i) and (ii) discussed above: some are known by the participants concerned to have been quite recent adoptions or name-changes; some are suspected to involve 18th century "neighbours" in USA, while others have surnames that imply the "event" occurred before the 17th century migration from Scotland.

While most administrators recognise the possibility of some participants in their projects having had NPEs in their ancestries, I suspect the extent of this feature is usually underestimated, in part due to the difficulties in

[49] In such projects only the TMRCA or paper trails can determine whether the cluster originally bore the surname or there has been an i-NPE relatively recently. This feature may contribute to why NPE rates appear lower that theoreticians assume.

[50] Strangely this project does not yet include any 18th or 19th century category (ii) i-NPEs, presumably due to low penetration.

[51] All have at least 31 markers, are in the largest cluster, and have TiPs with the modal signature of 98% or more.

identifying some i-NPEs, in part due to ignorance of empirical data on false paternity rates, and in part due to the difficulty and apprehension in expressing these complex and sensitive matters in a user-friendly way. Some of these issues are addressed in Appendix D, in which it is suggested that in most projects we should expect at least a quarter of all y-DNA test participants to have a NPE in their paternal ancestry.

8. **Features of clusters and singletons** (Appendix A, lines 35-38)

Having established some understanding of the concepts of clusters, singletons and the handling of NPEs, we can at last explore how these features emerge in our selected projects. The proportion of participants in each project that have been categorised into clusters ranges from 35% to 89%:[52]

% in clusters	< 50% of participants	50%-80% of participants	>80% of participants
	Taylor	Creer, Plant, Cruwys, Phillips, Wright, Walker, Williams	Dalton, Blair, Irwin

Why this diversity should be so great is unclear, although it appears that trade-name surnames are more difficult to categorise than place-name surnames. Counter-intuitively the differing definitions of "cluster" probably only make a small contribution to this diversity; nor do surname size, penetration or project bias seem to be relevant.

The size of the largest single cluster ranges also spans a wide range, from just 2% of participants to 66%:

Largest cluster	< 20% of participants	20%-60% of participants	>60% of participants
	Blair, Phillips, Wright, Walker, Taylor, Williams	Plant, Cruwys, Dalton	Creer, Irwin

It would seem that there is an inverse relationship between cluster size and surname size. The large size of the largest Irwin cluster suggests that project bias may also be relevant, although if so it is not clear why.

Of the ten projects so analysed, the size of the cluster that includes the earliest known ancestor range from 1% (Irwin, Taylor, Williams) to 64% (Creer) of all the participants.

8.1 Identification of clusters (Appendix A, line 40)

How project administrators choose to label their clusters may not seem an important point, but it is nonetheless revealing. Some make no identification apart from arbitrary numbering of the clusters (Blair, Phillips, Walker, Taylor, Williams). Some identify clusters by haplogroup (Dalton, Wright), by earliest ancestor (Plant, Wright, Williams), by geographic origin (Creer, Pomeroy, Dalton, Irwin) or by surname spelling (Plant, Cruwys), either alone or in conjunction.

8.2 **Sequence of participants within clusters** (Appendix A, line 44):

Similarly there are interesting variations of how participants' individual results are sorted within each cluster: some are sequenced by kit number (Plant, Dalton, Walker and Taylor), some by the marker counts (Phillips, Wright and Williams), and one (Irwin) by TiP probabilities from the modal signature.

The variations of individual administrators' choices in sections 8.1 and 8.2 above reflect differences in both data and attitude.

9. Other tools used for interpreting y-DNA test results

The survey identified several other tools available to project administrators that they may choose to use or disregard. The reasons for such choices are not always clear, but prima facie the evidence is apparently thus:[53]

9.1 Haplogroups (Appendix A, line 47)

Although haplogroups are primarily a "deep ancestry" feature, many project administrators recognise the slow mutation rates of their determining markers are a crude tool for assigning test results into clusters. Haplogroups are used as the prime tool for identifying clusters and assignment thereto in the Taylor project. Haplogroup details are not available for the Creer, Pomery and Plant projects. In the other projects it is unclear to what extent their inclusion is as a check on cluster assignments made, or as a crude criterion of "closeness", or simply to satisfy the interest in haplogroups of some of the project's participants.

[52] And conversely singletons range from 11% to 65% of all participants.

[53] SNPs, advanced markers, and RecLOHs are not considered in this paper as they figured little, if at all, in the survey.

Of the nine selected projects publishing haplogroup data, all have R1b1 dominant, with proportions ranging from 47% (Cruwys) to 100% (Creer) of their projects' participants.

9.2 Fast moving markers and rare marker values (Appendix A, lines 48 & 49)

Given the importance of mutation rates in many aspects of DNA test analyses there is rightly a keen interest in improving the understanding of this subject, but as yet little consensus and, I fear, little prospect that improved knowledge would relate directly to surname projects. The significance of fast moving markers and of rare marker values is deprecated by some project administrators but of interest to others, who believe these tools can occasionally be useful for identifying sub-clusters or individual participants who may be genealogically related.

Fast moving markers, as identified by FTDNA on their results pages, are monitored in the Creer, Pomeroy, Dalton, Blair, Irwin, Phillips, Wright, Walker, Taylor and Williams projects.

"Rare" marker values are monitored in the Blair, Irwin and Walker projects. There is no consensus on how rare markers are defined: I identify them as those with frequencies of less than 5% at www.ybase.org/statistics.[54]

9.3 TMRCA

Several tools enable the calculation of the time since the most recent common ancestor of two participants.[55] All require the resolution of the tests of the two participants concerned to be the same. Most assume some uniform mutation rate(s), though McGee allows a choice of mutation rates; FTDNA's TiP facility incorporates different mutation rates for each marker.

In adopting FTDNA advice on interpreting genetic distances, or using their TiP tool, all projects are relying on the results of TMRCA calculations. But as the graphs and tables summarising these calculations only represent average TMRCAs, and the spread of probabilities even within the timescale of inherited surnames is very wide, few administrators develop the application of this tool. The Dalton and Blair projects are exceptions.

9.4 Genetic distance matrices, including McGee's Genetic Distance calculator (Appendix A, line 50)

The concept of using probability matrices should, I believe, be seen in a wider context than its mechanics. The use of a probability matrix to assess the relationships between all participants in a project certainly has a mathematical justification, but I believe its application to surname projects lacks logic. The ethos of surname studies is that participants can be grouped into clusters whose members are all probably descended from a common ancestor. Unless identified by triangulation, the closest available estimate of the digital signature of that ancestor is the modal signature of the cluster. As we have already seen, some rule of thumb is necessary to decide whether there is a close match with this modal signature. The signature of each new participant can thus be compared with the relevant modal signature(s), which failing against each of the project's singletons, the latter leading to either the creation of a new cluster or to an additional singleton. Conceptually I thus see no benefit in assessing the "closeness" of every pair of participants in a project, or even within a cluster, although I do accept that genetic distance matrices may occasionally flag possible genealogical relationships that for various reasons may not be apparent in the relatively simple task of assigning participants to their appropriate cluster.

Dean McGee's program[56] creates a probability matrix in terms of genetic distance. Administrators may find this a convenient tool for small projects or clusters, but it becomes unwieldy for large clusters, and can only be used for a common resolution (typically 37 markers).

Of the selected projects it is apparently only used by the Creer, Dalton, Blair and Wright projects.

9.5 TiP (Appendix A, line 51)

The use of FTDNA's TiP (Time Indicator Projector) tool to define cluster membership in the Irwin project is summarised in Appendix C. This project also uses TiPs to sequence participants within each cluster (see section 8.2 above) and, for singletons, to identify their genetically closest participant.

The Wright project lists some TiPs. This tool is also used, albeit less obviously, in the Pomeroy, Plant, Cruwys, Dalton, Blair and Phillips projects.

9.6 Cladograms / Phylogenetic network diagrams (Appendix A, line 52)

[54] A more exhaustive list by Leo Little is at http://freepages.genealogy.rootsweb.ancestry.com/~geneticgenealogy/yfreq.htm.

[55] e.g. http://nitro.biosci.arizona.edu/ftdna/TMRCA.html; http://dna-project.clan-donald-usa.org/tmrca.htm.

[56] http://www.mymcgee.com/tools/yutility.html.

This tool uses engineering techniques to generate evolutionary trees to illustrate the most probable lines of descent of the signatures within a cluster.[57] Common resolution is necessary, but large clusters can be handled and the weighting given to assumed mutation rates can be varied. These diagrams are visually more digestible than matrices of genetic distances, but they are more difficult to create.

Of the selected projects the diagrams are only used in the Creer and Wright projects, but being explored in the Irwin project.

9.7 Triangulation

This term is used to describe the DNA testing of a third participant when two participants share a common pedigree but do not have close matching DNA signatures, in order to help establish the DNA signature of the common ancestor.[58] Its application is not restricted to lengthy pedigrees, although with these it can, at least in theory, identify the signature of the earliest common ancestor.

Triangulation is used in the Pomery, Cruwys, Irwin, Wright and Williams projects.

10. Single, plural and multiple origin families

For some projects, notably Creer, Pomeroy, Blair and Irwin, one of the principal goals is the use of DNA to determine whether the surname had a single ancestor (single-origin), a few ancestors (plural-origin), or many ancestors (multi-origin).

To date DNA has not shown convincingly that any of the selected projects are single-origin, but has thrown much light on this issue for the Creer, Pomeroy, Plant, Cruwys and Irwin surnames. On the other hand DNA has demonstrated that the Dalton, Blair, Phillips, Wright, Walker, Taylor and Williams surnames are, as expected, multi-origin.

Most projects recognise that participants with their surnames spelt similarly rather than identically can be genetically related, and that the spelling of surnames today is an unreliable indicator of cluster membership. On the other hand some such as Creer, Pomeroy, Plant, Cruwys and Irwin also recognise that surname spellings long ago, although unreliable, were sometimes indicative of a particular branch of a family.

Conclusions

It is clear that a survey of just 12 out of about 6,000 DNA surname projects cannot be considered representative. Nevertheless the concept of such a comparative study is shown to be justified, and several interesting issues emerge.

While some idealists may have dreamed that y-DNA tests would correlate closely with genealogy and pedigrees, there is now widespread if tacit recognition that, generally speaking, DNA surname projects are not yet achieving this idyll, except perhaps for a few of the smaller surnames. On the other hand they are certainly providing some very significant and tangible revelations to genealogists, including a structured statistical background to one-name studies, distinct "clustering" of participants with common ancestry, some clarification of geographical origins, the significance of NPE ancestries, assistance in resolving uncertainties in early pedigrees, and even sometimes identifying previously unknown distant cousins.

The selected projects encompass a wide diversity in content and presentational styles, and this survey has shown a wide range in terminology that is adopted and in the features that can be compared on a quantitative basis.

Perhaps the most significant conclusion of this survey is the extent to which the administration of small surname projects necessarily differs from that of large surname projects.

It has been shown that the advertised sizes of surname projects can be misleading, and even the number of tests completed is of limited value. An alternative measure is developed, the "penetration" of individual projects, defined as the number of test results expressed as a percentage of the world population of the same surname. Penetrations of the selected projects are typically 0.06%, but range from 0.02% to 1.5%. There appears to be some inverse relationship between penetration and surname size.

The majority of the populations of the selected projects are resident in the New world (between 68% and 86%), as are the places of residence of participants in these projects (between 64% and 94%, with one exception). When determinable, the ratio of New world participants to New world population is generally close, i.e. most projects show little geographical "bias" in representing the actual diaspora of their surname.

The survey has identified a wide range in the rules of thumb used to determine a "close match", even if this may not make much difference in practice. Most are

[57] For methodology see http://www.fluxus-engineering.com/sharenet.htm, and as an example of application see http://www.ewingfamilyassociation.org/DNA_Project/index_Y-DNA.html > Results Directory > About Diagrams.

[58] For a fuller discussion see http://www.kerchner.com/triangulation.htm.

based on genetic distance, and thus make no allowance for the differing average mutation rates of individual markers.

As an alternative rule of thumb a case is made for the use of the TiP tool, for example a 80% probability criterion at 24 generations, though it is accepted that this may not suit all project administrators.

All the selected projects seek to assign their participants into clusters, and some use the modal signature of these clusters as the basis for listing and comparing individual test results. The ability to assign test results into clusters, as opposed to residual singletons, shows a wide diversity, from 35% to 89%. The reasons for this diversity remain unclear, but may be related to surname type. The number of participants in the largest cluster in each project range from 2% to 66%. These ratios inversely reflect surname size and any project bias.

Some of the issues surrounding NPE's are identified, and several ideas are developed to help with the handling of this sensitive but probably under-appreciated aspect of surname projects.

This study suggests that the most useful tools for analysing y-DNA test results are the identification of clusters and cluster modal signatures, FTDNA's TiP tool, possibly cladograms and triangulation, and of course paper trails. Other tools such as haplogroups, fast moving and rare markers, TMRCAs, genetic distances and genetic distance matrices, though by no means irrelevant to surname studies, can be a distraction. Some of these are relics of early studies, and today may have more relevance to deep ancestry research.

There does seem to be a pressing need for proposals for the standardisation in the terminology used in y-DNA surname projects, and "best practices" in their administration, but this is not the occasion to develop this process. Nor am I suggesting that volunteer administrators should necessarily change their established practices!

No doubt some of the present constraints and uncertainties associated with current y-DNA projects may be overtaken by advances in testing techniques. Nevertheless there is clearly scope for benefits to accrue from increasing the penetration of surname projects large and small, and from more rigour in the assignment of participants to clusters, and from greater recognition of the NPE phenomena. These developments are likely to lead to a lowering of the proportions of singletons, to more reconciliations of genetic and genealogical data, both pre- and post-migration, and eventually to the challenging some of the assumptions in surname dictionaries.

Acknowledgements

I am grateful to the project administrators of my selected projects, all of whom have kindly given their permission to quote their work and many of whom, and particularly Debbie Kennett, have offered useful comments on drafts of this paper. But of course all the opinions expressed, and remaining errors, are my own.

Appendix A:
Summary of analyses the 12 selected projects

1	Surname	Creer	Pomeroy	Plant	Cruwys	Dalton	Blair	Irwin	Phillips	Wright	Walker	Taylor	Williams
2	traditional type	place	place	place/nick	place	place	place	place	personal	trade	trade?	trade	personal
3	traditional origin	I of Man	England	England	England	England	Scotland	Scotland	England	-	UK	England	-
	Project size												
4	Advertised size	28	75	41	55	132	169	205	439	336	620	348	751
5	Year/month founded	2005/6	2000	2001	2007/9	2003/5	2002/6	2005/10	2006	2002/8	2001/5	2004	2003/1
6	av. growth rate	6 p.a.	8 p.a.	5 p.a.	22 p.a.	19 p.a.	22 p.a.	45 p.a.	110 p.a.	45 p.a.	70 p.a.	60 p.a.	105 p.a.
7	Funding available?	no?	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	yes
8	yDNA tests completed	26	118	37	46	126	157	188	456	302	548	306	747
9	completed/advertised	0.93	1.57	0.90	0.84	0.95	0.93	0.92	1.04	0.90	0.88	0.88	0.99
	Population												
10	Spelling variants used	1	7	10	26	5	9	35	12	10	1	6	1
11	Approx world population	1,759*	19,462	58,691	71,426	104,804	131,224	278,807	755,367	793,725	860,856	1,763,288	2,261,019
12	suggested category	very small	small	medium	medium	medium	medium	medium	large	large	large	large	large
13	penetration	1.48%	0.61%	0.06%	0.06%	0.12%	0.12%	0.07%	0.06%	0.04%	0.06%	0.02%**	0.03%
	Old/New world ratios												
14	Population drift	70%	85%	68%	84%	81%	86%	82%	80%	77%	77%	79%	80%
15	Project drift		c26%	64%	83%	82%	94%	87%	93%				
16	Project bias		c0.30	0.94	0.99	1.01	1.09	1.06	1.16				
	Paper trail data of participants' earliest ancestors												
17	Details listed	pedigrees	pedigrees	earliest	pedigrees	pedigrees	earliest	earliest	pedigrees	pedigrees	pedigrees	pedigrees	pedigrees
18	England & Wales	46%	81%	56%			5%	2%	10%		5%	9%	) 5%
19	Scotland	4%	0%	0%			18%	25%	1%		3%	2%	)
20	Ireland	0%	5%	5%			29%	35%	2%		2%	2%	)
21	other Old world	19%	3%	13%			1%	2%	2%		1%	0%	)
22	(mainly in)	(IofM)	(Russia)	(France)			(Neth)	(Germany)	(Poland)		(Germany)	-	
23	New world only	42%	11%	23%			45%	28%	73%		67%	28%	
24	no data available	0%	0%	3%			2%	8%	12%		22%	59%	
25	pre 1599		6%	2%				6%	2%		1%	2%	)
26	1600-1799		56%	46%				47%	48%		42%	24%	) c25%
27	post 1800		38%	39%				38%	35%		25%	21%	
28	no data available		0%	12%				11%	15%		32%	53%	
	Resolution												
29	12 markers	0%	3%	23%	0%	2%	4%	10%	17%	16%	21%	32%	19%
30	25 markers	62%	3%	31%	0%	22%	31%	2%	8%	10%	16%	3%	17%
31	37 markers	38%	30%	43%	76%	37%	50%	50%	45%	44%	41%	31%	39%
32	67 markers	0%	7%	0%	24%	39%	16%	35%	19%	26%	23%	34%	23%
33	other	0%	57%	3%	0%	0%	0%	3%	12%	4%	0%	0%	2%
	Clusters												
34	modal signature			yes	no	yes	yes	yes	yes	some	no	some	yes
35	singletons	?36%		30%	33%	17%	11%	13%	25%	23%	22%	65%	28%
36	all clusters	?64%		70%	67%	83%	89%	87%	75%	77%	78%	35%	72%
37	largest cluster	64%		43%	28%	35%	14%	66%	5%	4%	7%	2%	3%
38	earliest cluster	64%		17%	9%	5%	13%	1%	5%		3%	1%	1%
39	no. of clusters	13	8	6	8	13	9	15	63	53	58	37	93
40	primary methods	location	location	spelling	h'group	h'group	ancestors	location	-	h'group	h'group	h'group	ancestors
41	of identification			ancestors	ancestors	location				ancestors			h'group
42				location	spelling					location			
43	no. identified origin	?	?5	all	all		2	all	0	few	0	0	0
44	sort within cluster			kit, no		kit no.		TiP	markers	markers	kit no.	kit no.	markers
	Non Paternal Events												
44	i-NPEs (other DNA)	yes	no		yes	yes	yes	8%	3%	yes	yes		yes
45	e-NPEs (other names)	no	no	no?	no	no?	no?	10%	3%	yes	yes	yes	yes
	Tools												
46	Halopgroups R1b1	100%	NA	NA	47%	90%	no	92%	53%	59%	59%	71%	76%
47	fast moving markers	yes	yes	no	no	yes	yes	yes	yes	yes	yes	yes	yes
48	rare markers	no	yes	no	no	yes	yes	yes	no	no	no	no	no
49	use of McGee	yes	no	no	no	no	yes	no	no	no	no	no	no
50	use of TiP	no	yes	yes	yes	yes	part	yes	yes	part	no	no	no
51	use of cladograms	yes	yes	no	no	yes	no	yes	no	no	no	no	no
	Earliest dates												
52	of surname	1511		1210	1200s	1153	1205	c1190	1273			c1200	
53	in earliest cluster		1066	1647	1449	1230	1547	1323	1540		c1500	1485	1575
54	in largest cluster	c1630		1647	1690	1650	1661	1484	1540	1665	1660	1715	1669
55	Surname type	single?	single?	plural?	plural?	multi	multi	plural?	multi	multi	multi	multi	multi

*: All Creers, not just Isle of Man; **: 0.03% if Ancestry project is included

Appendix B:
Calculation of world populations of surnames

In August 2008 University College London launched a website listing the frequencies of surname spellings on a geographical basis around the world (www.publicprofiler.org/worldnames).[59] This free-to-use facility provides detailed breakdowns of the frequencies of some eight million spelling variants of surnames in 25 countries in 2000-2005. This is apparently the best publicly available database for calculating approximate current national and world populations for different spellings of surnames,[60] but it is important to recognise its limitations. These include:

- the data cannot distinguish between surnames that are similarly spelt but wholly unrelated;
- the data has been derived from telephone directories and electoral rolls, and no details are available on how these figures have been processed to derive the published total population frequencies; the data is thus indicative rather than definitive;
- no surname data is included for Portugal, or for most of Central and South America, Africa and Asia;
- population frequencies for any single surname spelling are not available for more than the ten countries.

The latter two points give a bias towards the derived world surname populations being understated. Hence it is important to recognise the population figures derived are only approximate.[61] So although this database is not intended for scientific application, and its limitations are significant, it does give approximate population figures for a very wide variety of surname spellings, and accurate population data is not critical for the applications to which it is put in this paper.

The calculation of approximate world population of a surname using the UCL Worldsurnames database is straightforward, even if a little involved. For each surname spelling variant the website lists the FPM (frequency per million), by country. To convert these frequencies for each country into population figures requires knowledge of the population of each of these countries. The CIA's World Factbook is a convenient source.[62] Multiplying the UCL frequencies by today's population of each country gives the total population for each surname spelling, i.e.

Country population for each surname spelling = frequency per million for each spelling x country population in millions

If, as for most surname projects, there is more than one spelling variant, then these frequencies for each country have to be summed:

$$\text{Approx. country population for each surname} = \Sigma(\text{country populations for each surname spelling})$$

The approximate world population for each surname is the sum of the relevant country populations:

$$\text{Approx. world population for each surname} = \Sigma(\text{country populations for each surname})$$

A pro-forma Excel spreadsheet can be set up to "automate" this process, so that only the spellings and their frequencies need to be transcribed manually, and all the other data is printed automatically.[63]

In the attached example of the use of this pro-forma spreadsheet, for the surname Meates, the data entered manually is in italics – the remaining data has been calculated automatically. This example is not typical, for two reasons: first the number of surname spellings used, 19, is greater than that needed for most projects (see Appendix A, line 10); and second, the spelling "Mates" shows that this is version is particularly prevalent in Hungary and India, completely distorting the distribution of the anglo-saxon surname. Fortunately this example of "overlapping" of quite different surnames is unusual. Except for the Creer, Plant and Cruwys projects, which recognised this dimension from their inception, this problem was not significant with any of the other nine projects selected for this paper.

As shown in section 2.2 above, the calculation of the penetration of a surname project is simply the ratio of the number of completed DNA tests to the world population:

$$\text{Penetration \% for each surname} = \frac{\text{No.of completed yDNA tests for the surname x 100}}{\text{World population of the surname today}}$$

[59] It should be recognised this database is independent of UCL's analyses of the origins of names.

[60] A similar tool is http://www.dynastree.co.uk/maps/detail/dillman.html but this only covers nine countries, includes a much smaller range of surname spellings, and is based solely on telephone directory data, which with the advent of market competition and mobile phones now represent a smaller proportion of the population than the Worldnames facility (e.g. 23% vs. c.40% for UK). However the Dynastree website does have some presentational advantages which may make it more attractive for some purposes.

[61] Even so the global population derived for the Pomeroy surnames is 40% more than that estimated by Chris Pomery, although Pomery admitted his own figure was a very rough estimate (pers. comm. 2009).

[62] Obtained from https://www.cia.gov/library/publications/the-world-factbook/index.html. The latest available data is for mid 2009; this use of more up-to-date data than that used by UCL is not an issue.

[63] A copy of this pro-forma Excel spreadsheet is available for download at http://www.jogg.info//62/files/SurnameMatrix.xls.

Population data for surnames like MEATES

(1) Population density data: For each chosen surname spelling, enter frequencies per million from www.publicprofiler.org/worldnames

Surname	Old world, frequencies per million																	New world, frequencies per million							Surname
spelling	UK	Austria	Belgium	Denmark	France	Germany	Hungary	Ireland	Italy	Luxemburg	Netherl'ds	Norway	Poland	Serbia	Slovakia	Spain	Sweden	Switzerl'd	Argentina	Australia	Canada	India	New Zeal'd	USA		spelling
Mate	*0.00*	*13.10*					*1333.6*							*31.09*	*72.50*	*93.77*	*6.32*			*21.88*	*16.50*	*37.67*	*20.10*		*some < 6.32*	Mates
Mates	*5.08*	*1.98*					*14.15*	*24.00*						*1.17*	*2.90*	*10.06*		*1.28*			*1.38*			*3.44*	*some < 1.17*	Mates
Matt	*2.21*	*175.2*			*10.6*	*52.26*	*14.15*	*1.37*									*3.79*	*46.64*			*23.38*			*15.41*	*some < 1.37*	Matt
Matte	*0.00*		*14.47*		*14.15*	*4.29*			*1.70*			*0.85*					*2.53*		*0.93*		*140.28*	*3.77*		*9.53*	*some < 0.85*	Matte
Meat	*0.02*		*1.72*		*0.69*						*0.43*													*0.08*		Meat
Meate	*0.04*																									Meate
Meats	*3.2*				*0.44*	*0.04*																		*0.75*		Meats
Meates	*1.34*							*16.46*															*21.16*	*0.01*		Meates
Meitts																									*nil*	Meitts
Meot			*0.29*		*0.89*	*0.04*					*9.31*										*0.46*			*0.11*		Meot
Meote																									*nil*	Meote
Miat	*0.02*				*0.84*	*0.14*			*0.38*			*0.28*												*0.01*		Miat
Miot	*0.07*		*15.33*		*38.71*	*0.33*		*1.03*	*5.02*									*1.28*	*0.60*		*0.69*			*0.21*	*some < 0.07*	Miot
Mayte	*0.02*				*0.54*	*0.02*					*0.21*											*0.27*		*0.01*		Mayte
Mayett									*0.06*															*0.06*		Mayett
Mayot	*0.20*		*2.01*		*15.14*													*1.92*	*0.04*							Mayot
Mayott	*0.09*																				*0.23*			*0.45*		Mayott
Myatt	*56.79*		*0.29*	*0.98*	*0.15*			*1.71*										*0.64*		*25.89*	*35.3*		*8.92*	*12.52*	*some < 0.15*	Myatt
Myott	*1.16*																				*0.46*			*1.43*		Myott
																										0
National pop., m	62.0	8.4	10.8	5.5	65.4	81.8	10.0	4.5	60.2	0.5	6.6	4.9	38.2	7.8	5.4	46.0	9.3	8.8	40.1	22.2	34.0	1178.7	4.4	308.9	2024.4	

(2) APPROX. TOTAL POPULATIONS FOR EACH CHOSEN SPELLING: These figures are calculated automatically by mutliplying population density data by national population.

Surname	Old world, approx. total population																	New world, approx. total population						Approx.tot.	%	Surname
spelling	UK	Austria	Belgium	Denmark	France	Germany	Hungary	Ireland	Italy	Luxemburg	Netherl'ds	Norway	Poland	Serbia	Slovakia	Spain	Sweden	Switzerl'd	Argentina	Australia	Canada	India	New Zeal'd	USA	population		spelling
Mate	0	110	0	0	0	0	13336	0	0	0	0	0	0	243	392	4313	59	0	0	486	561	44402	88	0	63989	59%	Mates
Mates	315	17	0	0	0	0	142	108	0	0	0	0	0	9	16	463	0	11	0	0	47	0	0	1063	2189	2%	Mates
Matt	137	1472	0	0	693	4275	142	6	0	0	0	0	0	0	0	0	35	410	0	0	795	0	0	4760	12725	12%	Matt
Matte	0	0	0	0	925	351	0	0	102	0	0	4	0	0	0	0	24	0	37	0	4770	4444	0	2944	13601	13%	Matte
Meat	1	0	0	0	45	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	25	74	0%	Meat
Meate	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0%	Meate
Meats	198	0	0	0	29	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	232	462	0%	Meats
Meats	198	0	0	0	29	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	232	462	0%	Meats
Meates	83	0	0	0	0	0	0	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	93	3	253	0%	Meates
Meitts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	Meitts
Meot	0	0	0	0	58	3	0	0	0	0	61	0	0	0	0	0	0	0	0	0	16	0	0	34	173	0%	Meot
Meote	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	Meote
Miat	1	0	0	0	55	11	0	0	23	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	95	0%	Miat
Miot	4	0	0	0	2532	27	0	5	302	0	0	0	0	0	0	0	0	11	24	0	23	0	0	65	2993	3%	Miot
Mayte	1	0	0	0	35	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	318	0	3	361	0%	Mayte
Mayett	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	22	0%	Mayett
Mayot	12	0	0	0	990	0	0	0	0	0	0	0	0	0	0	0	0	17	2	0	0	0	0	0	1021	1%	Mayot
Mayott	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	139	152	0%	Mayott
Myatt	3521	0	0	5	10	0	0	8	0	0	0	0	0	0	0	0	0	6	0	575	1200	0	39	3867	9231	9%	Myatt
Myott	72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	442	529	0%	Myott
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0%	0
Total number	**4553**	**1598**	**0**	**5**	**5401**	**4676**	**13619**	**201**	**431**	**0**	**66**	**6**	**0**	**252**	**407**	**4776**	**118**	**455**	**63**	**1060**	**7435**	**49164**	**221**	**13829**	**108336**		
%	**4.2%**	**1.5%**	**0.0%**	**0.0%**	**5.0%**	**4.3%**	**12.6%**	**0.2%**	**0.4%**	**0.0%**	**0.1%**	**0.0%**	**0.0%**	**0.2%**	**0.4%**	**4.4%**	**0.1%**	**0.4%**	**0.1%**	**1.0%**	**6.9%**	**45.4%**	**0.2%**	**12.8%**		**100%**	
								Old world total = 34%													**New world total = 66%**						

NB Numbers of households and numbers of adult males are about 40% of total population figures (www.nationmaster.com > People Statistics > Average size of households)

Appendix C:
Use of the TiP tool for defining "close matches" and as a criterion for cluster assignment

As the administrator of a large surname project I often have to respond to questions such as:

- How reliable are 12-marker test results, and what benefits accrue by upgrading to 25, 37 or 67 markers?
- What, quantitatively, is meant by DNA results that are "near identical" or "close matches"?
- How valid are rules-of-thumb such as "1/12", "3/25" and "5/37" or "50% TiP" as criteria to address the question "Are participants A and B genetically related or not, within the era of hereditary surnames?"
- How can one compare test results with different resolutions (i.e. a mixture of 12, 25, 37 and 67 markers)?
- On what basis do you assign participants to a particular cluster? Why have you included A but not B?

Besides these general questions, as administrator of the Clan Irwin project I have some more specific challenges:

- We now have approaching 200 participants, necessitating rigour and consistency in interpreting test results.
- We have several clusters that are clearly genetically unrelated, and one of which is very dominant. We are also fortunate in being able to associate all our clusters with their probable "Old world" place of origin.
- Some of our participants have paper trails going back as early as the 14th century.
- It is difficult to draw general conclusions when our main cluster includes two full brothers who inherit a pedigree of 11 generations but have a mismatch of 2/25, and also 16 other participants with mismatches of 0/67 or 1/67 but who have no apparent genealogical relationship within, typically, at least 8 generations.[64]
- Although I appreciate that marker mutation rates vary considerably, I lack detailed understanding of why this is so, or of how such knowledge might help administrators, and I remain apprehensive whether such knowledge could answer my initial questions above.[65]

I have thus sought some single, simple, robust, transparent and repeatable "rule of thumb" with which to decide whether or not two participants can be considered genetically related within the timeframe relevant to surname projects. I considered tools such as genetic distance, probability matrices such as McGee's calculator, and even full exploitation of FTDNA's TiP tool,[66] with its multi-generation probabilities and paper-trail recalculation facility. These all have some value, but none provide the tool I sought.[67]

However within the mass of data available from the TiP facility I suspected there might be some feature which met my needs, especially as uniquely this tool takes account of the non-uniform mutation rates of individual markers. The TiP facility is a sophisticated aid, and the application of its full potential is not straightforward, especially for large projects that were already established when it was first introduced in 2004. Furthermore, for commercial reasons its derivation remains confidential,[68] while its exposition to two decimal places gives a misleading impression of reliability. For these reasons it receives some legitimate criticism and it is not surprising that some cognoscenti view it with disdain, while many newbies view it with bewilderment. It also cannot be applied to non-FTDNA data. But despite these disadvantages it is the only tool available to project administrators that takes account of test results with different resolutions, of the different average mutation rates of individual markers, and, apparently, of the rare instances of null makers and RecLOH events.[69]

The question thus arises of whether it is possible to select from the sophisticated TiP facility some simple quantitative yardstick that can provide a simple "Yes/No" answer to the question "Are participants A and B genetically related / 'a close match' "? Two simplifications of the TiP facility quickly become apparent:

- for this application it is appropriate, as well as convenient, to dispense with FTDNA's "paper trail" refinement, and anyway the "no paper trail" default is a "worst case";
- the 24-generation TiP probability is likewise a worst case,[70] going back to the earliest use of surnames without unnecessarily invoking "deep ancestry" considerations;

[64] See www.clanirwin.org > DNA Study.

[65] I suspected this frustration was a personal problem until at the GOONS Seminar on "DNA Developments" in February 2010 it became apparent that administrators of other surname projects were posing similar questions, including those with much more knowledge of mutation rates than myself.

[66] A description of FTDNA's TiP is given at https://www.familytreedna.com/faq-tip.aspx.

[67] I accept, of course, that all yes/no tests are only as good as the data available at the time, and that DNA test results are probabilistic by nature, and so no such tests can be considered infallible.

[68] Intuitively, as a genealogist, I suspect this confidentiality, however justifiable, could mask some bias in the TiP algorithms toward probabilities greater than rigorous peer review might allow. I also gather some geneticists suspect the mutation rates used may be too high. But such bias, even if true, is irrelevant if TiPs are only used, as they are in this paper, in a relative context.

[69] See http://en.wikipedia.org/wiki/RecLOH.

[70] I justify this for Scottish surnames in Appendix D above. For English and Irish surnames it may not be a worst case, but in practice for the purpose of a yes/no test of "closeness" this nicety is not significant. A disadvantage of adopting the 24-generation TiP is that many participants within a project will

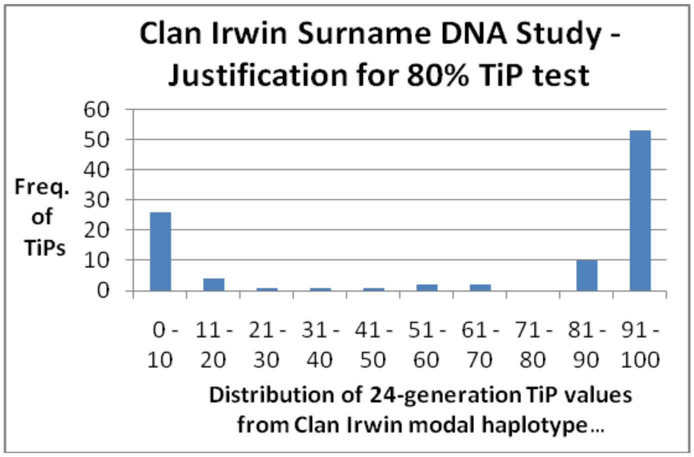
- using this 24-generation TMRCA as an input parameter, and the resulting probabilities as output.

On this basis the use of a 50% 24-generation, "no paper trail" TiP test for closeness cannot be a worse "rule of thumb" than the various other "rules of thumb" that have been identified in section 7 above, even if, like all rules of thumb, it cannot exclude all false positives or include all valid linkages.

For the Irwin project (see www.clanirwin.org > DNA Study, and summarised in Appendix E below) I have taken several further steps:

- moving from a probability criterion of 50% to 80%. This I have done for greater confidence. In practice this refinement has only meant classifying an additional 6 (out of 188) participants as singletons, and of these 4 had only 12-marker resolution tests.[71]
- using FTDNA's GAP Report pages to facilitate comparison of each new participant with the participant(s) sharing the modal signature of each cluster, and hence assign them to the appropriate cluster.
- also noting for each participant his haplogroup and genetic distance at 12 and, where relevant, 25, 37 and 67 markers, in order to demonstrate the consistency of his TiP probability with this data.
- sorting and listing the participants within each cluster in order of their TiP probabilities.
- noting for each singleton details of his "nearest match", using the same process.

The following histogram shows the frequencies of 24-generation TiPs for participants by decile, justifying the selection of 80% as an arbitrary but robust criterion.[72]



These procedures have the advantages of:

- being able to list and compare on a rigorous basis participants' test results with varying resolutions and varying numbers of generations since their earliest confirmed ancestor;
- showing clearly the limitations of relying on low resolution tests: it is apparent that as the resolution increases so the associated TiP with the cluster modal signature tends towards 0% or 100%;
- illustrating the limitations of relying on genetic distance alone: these cannot illustrate the probabilistic dimension, and also fail to take account of the very different average mutation rates for individual markers;
- providing a simple and transparent criterion which includes explicit justification for why one participant qualifies for assignment to a cluster when another does not;

In the Irwin project the arbitrary 80% criterion excludes some participants with 0/12 and 1/12 mismatches but "allows" some with 3/12 mismatches; it excludes some with 4/25 mismatches but allows some with 5/25 mismatches; and it excludes some with 8/37 mismatches but allows some with 6/37 mismatches - indicative of the different average mutation rates adopted within the TiP alogorithm.[73]

have a TiP with the relevant modal signature of over 99%. This can be clarified by using the otherwise misleading two decimal places of the TiP probability, and/or including a secondary indicator for such participants of, say, 8-generation TiPs.

[71] I could have opted for 90% or 95% probabilities, but this would reduce many more participants to singleton status without any apparent justification.

[72] This histogram reflects the unusual characteristic of this project in two thirds of participants being members of a single genetic family. For other surname projects the 80% criterion might be less clear, but the principle would still apply.

[73] Most 12-marker tests generate TiPs above the 80% criterion, clearly indicating higher resolution is desirable. A few 12- marker tests do meet the 80% criterion, and I accept that higher resolution testing could show that these participants after all do not belong to the cluster first indicated.

While I find this 80% 24-generation TiP with the cluster modal signature to be a most convenient and reliable rule-of-thumb for determining whether two participants are genetically related or not within the surname era, and for assigning a participant to a particular cluster, I accept that other administrators may find it less useful. Nor do I see it as a panacea: not only may occasional exceptions be necessary, but the sorting of participants' test results into clusters is only one stage in the process of analysing y-DNA data. Other stages include:

- seeking means of sub-dividing each cluster/genetic family into branches, on the basis of fast-mutation markers, rare marker values, cladograms/phylogenetic network diagrams, FTDNA's "unique haplotype" pages, triangulation, paper trail data etc.;
- identifying and explaining the likely geographic origin of each cluster and branch;
- comparing and developing the relevant genealogical data available on the individual participants in each cluster and branch, to seek possible genealogical connections;
- refining and pursuing other goals of the project.

Appendix D:
Quantification of rates of Non Paternity Events

There is little consensus on typical historical false paternity rates. Laslett et al. recorded bastardy rates of 1-7% per generation, averaging 3-4%, in 98 English parishes from 1540 to 1900, and 4-11% in Scotland in the mid 19th century.[74] McEvoy and Bradley calculated one Irish family had 1.6% per generation.[75] King and Jobling demonstrated that historically false paternity rates probably lie between 1% and 4.5% per generation, and adopted 2% for their simulation modelling.[76] FTDNA suggest using between 1% and 2% per generation.[77]

But all these "per generation" rates are cumulative. Plant suggested that "one can theoretically expect that around half of randomly selected, modern bearers of a populous, single family surname will remain free of ancestral introgressions."[78] Does this infer that at least in a single-name surname DNA project about half of the volunteer participants (as opposed to those proactively recruited because of their known pedigrees) should be expected to be NPEs?[79]

Plant has developed the arguments further. He suggested the probability P of a participant being a biological descendant of his surname's progenitor can be approximated from the formula P% = $(1\text{-}p)^n$x100, where p is the fractional probability of a false paternity per generation and n is the number of intervening generations:

	p = 1%	p = 2%	p = 5%	p = 10%	p = 30%
n = 5	95%	90%	77%	59%	17%
n = 15	86%	74%	46%	21%	0%
n = 25	78%	60%	28%	7%	0%
n = 35	70%	49%	17%	3%	0%

From this I deduce that the probability of participants not matching the modal signature of a single-origin surname is (1-P)%. Furthermore, the number of intervening generations may be calculated from n = (T/t + 1), where t is the average generation interval in years (see section 3 above), and T is the number of years since the earliest hereditary holder of the name. Simplistically let us assume that hereditary surnames were first found in Ireland in the 10th century, in England in the mid 11th century, and in the Lowlands of Scotland in the late 13th century. Combining these elements, and adopting FTDNA's guide for non-paternity rates of 1-2% per generation, gives the following cumulative probabilities that may be expected for the proportion of participants in a single name project to have NPE ancestry:

Date of earliest ancestor with hereditary surname		T	t		
			25 yrs/generation	33 yrs/generation	40 yrs/generation
	AD1700	300 years	12-22%	10-18%	8-17%
	AD1500	500 years	18-35%	15-28%	12-24%
Scotland	AD1300	700 years	25-46%	19-36%	17-32%
England	AD1100	900 years	33-55%	25-44%	20-37%
Ireland	AD 900	1100 years	46-65%	31-52%	25-43%

These percentages are probably also indicative of the considerable proportion of NPEs that are to be expected to be embedded within the y-DNA test results for multi-origin surnames.[80] And of course per-generation false paternity rates higher than 2% are quite possible: Plant quoted an example of 30%, and we should not forget that in the UK today 50% of the current generation of children are born out of wedlock!

So even assuming a most conservative false paternity rate of 1% per generation, in the Irwin project where I believe t = 35 years and T = 700 years, then on the above basis I should be expecting to find at least about 20% of our participants have

[74] Peter Laslett et al (editors) 1980 *Bastardy and its Comparative History.*
[75] Plant 2009, 8.
[76] King and Jobling 2009, 1095.
[77] FTDNA website FAQ id 567.
[78] Plant 2009, 3. Plant's comment was no doubt based on Bryan Sykes's finding that only 44% of his participants matched the modal haplotype, and Sykes's study was in turn based on only four markers.
[79] The lack of appreciation of the likely frequency of NPE's was illustrated by the gasps of disbelief when at the GOONS seminar in February 2010 I ventured to suggest that perhaps a quarter of those present had a NPE in their paternal ancestry!
[80] I justify this broadening of the argument on the basis that participants in a multi-origin surname project represent the sum of a number of single-origin surnames, and I find it difficult to assume that multi-origin surname holders behaved significantly differently to single-name surname holders.

NPEs in their paternal ancestries. In fact to date I have only identified 16%, and that is including both i-NPEs and e-NPEs, although I suspect the latter should be excluded from this count![81]

One reason why relatively few i-NPEs are identified is the difficulty in recognising them as such in a multi- or plural-origin surname project (see section 7 above). In other words, in such projects, some of the clusters may include i-NPEs even though no evidence survives to identify them as such.

[81] As one project's e-NPE is another project's i-NPE, perhaps e-NPEs should only be counted in the projects of their "new" surnames, and including their number here is double counting.

Appendix E:
Overview of the Irwin Surname Project[82]

This project’s 188 test results have been categorised on the basis of a 80% 24-generation TiP thus:[83]

Cluster identifier	Geographical origin		i-NPE surname	Haplo-group	No. of participants total	resident in Old world	Earliest confirmed ancestor	Most common Old world Spelling
BA	Scotland	Borders		R1b1	105	10	1484	Irving, Urwin
”		”	e-NPE	R1b1	15	2	1660	various
NB		Bell	i-NPE	R1b1	5	0	1755	Irving
NC		Carruthers	”	I1	1	0	1791	”
ND		Dodd	”	I1	2	0	1812	”
NE1		Elliot 1	”	I1	2	0	1765	”
NE2		” 2	”	R1b1	2	1	1738	”
NG		Graham	”	I1	1	0	1811	”
NJ		Johnston	”	R1b1	2	0	1750	”
DA		Aberdeenshire		R1b1	2	2	1323	Irvine
O1		Orkney 1		R1b1	2	1	1460	”
O2		” 2		R1b1	2	0	1598	”
PA		Perthshire		R1b1	2	2	1730	”
IL	Ireland	Leinster?		I	6	0	1725	Irwin (O’Hirewen)
IM		Munster		R1b1	3	1	1785	” (O’Ciarmhachain)
G	Germany /Netherlands			R1b1	6	0	1762	(Arwine)
Singletons		?		G	1	0	?	Irwin
”		Scotland		I	2	1	?	Irving
”		?		J2	1	0	?	Irwin
”		Ulster		R1a1	2	2	1830	Irvine
”		various		R1b1	17	2	1650	Various
Too few markers (12) for categorisation				R1b1	7	0	1757	Various

It can be seen that we have been able to associate all our clusters, (including all 7 i-NPE clusters)[84] with geographical origins in the Old world. The individual clusters are best reviewed from the bottom up:

The Germany/Netherlands cluster (all US-resident participants) was hitherto unsuspected. More work is required on this family, including finding potential participants still resident in Europe. While originally Arwine, the US-resident participants now include some Irwins, while the Borders cluster includes some Arwines - clear evidence of unstable surname spelling during early settlement in USA. Other clusters include similarly unstable spellings, many pre-dating migrations when surname misspelling was associated with Ellis Island.

The two Irish clusters are anglicised versions of gaelic names listed in surname dictionaries. The Munster cluster has clear evidence of Co.Tipperary roots, gaelic speech and catholic religion. More work is required to confirm the other cluster did in fact originate in Leinster. But it seems clear that both these clusters, like the Germany/ Netherlands cluster, never had any connection with Scotland. No Irwins originating in England have yet been identified: it is unclear whether this is due to poor penetration, or to an error in the surname dictionaries.

A tradition, first recorded in the 17th century, claims that within Scotland the surname was single-origin. Pro-active recruiting of participants resident in UK with lengthy and reliable pedigrees has revealed four clusters with clear geographic origins distinct from the Borders cluster, all of which are compatible with branches named in the tradition. Whether this development shows that the name in Scotland is in fact plural-origin, or that were several early i-NPEs, remains a contentious issue. Further participants with lengthy pedigrees are being sought.

[82] For full details see www.clanirwin.org > DNA Study.

[83] The criterion has been modified on the basis of the caveats to the definition in Appendix C above. The adoption of the 80% criterion rather than 50% only relegated 6 participants to singleton status, and of these 4 have only 12-marker tests.

[84] For explanation of the terms i- and e-NPEs see section 7 above.

Within the Borders cluster the few UK-resident participants include two lengthy and reliable pedigrees with Dumfriesshire origins. Of the many New world resident participants most claim Scots-Irish ancestry but few have pedigrees reaching back to Ireland, and only one back to Scotland. All have been gratified the project has been able to confirm their Scottish roots, although to date very few have been able to identify genealogical relationships with other participants. Hopefully current work on a cladogram of this large cluster will help, but clearly the cause is our low penetration (0.05%) of the c.170,000 Irwins etc. now residing in USA.

Seven small but distinct clusters of i-NPE Irwins have all been found to share the DNA signatures of other Borders families, also suggesting their "events" occurred before the 17th century plantation of Ulster.

Of the e-NPE participants who have joined the project because they closely match the Borders modal signature, some are aware why their name recently changed from Irwin, others suspect the "event" occurred in the 18th century, and some, having surnames of other Borders families, are suggestive of "events" that occurred before c.1600.