Journal of Genetic Genealogy

Journal: www.jogg.info
Originally Published: Volume 9, Number 1 (Fall 2021)
Reference Number: 91.007

Y-DNA SNP-BASED TMRCA CALCULATIONS FOR SURNAME PROJECT ADMINISTRATORS

Author(s): James M. Irvine

Journal of Genetic Genealogy

Y-DNA SNP-based TMRCA calculations for Surname Project Administrators

by James M Irvine, Administrator, Clan Irwin Surname DNA Project (jamesmirvine@hotmail.co.uk)[1]

Summary

The calculation of a "Time to Most Recent Common Ancestor" (TMRCA) using relevant SNP data can be so simple that many genealogists are tempted to use this tool and to draw inaccurate, imprecise and unwarranted conclusions, however unintentionally. Conversely the calculation of the Confidence Intervals (CIs) that should accompany such calculations is complex and rarely attempted. This paper is not promoting some new panacea, but draws in part on a novel analysis of 17 samples of SNP counts to help genealogists to understand why the popular use of SNP-based TMRCAs without CIs is misguided, why in practice these CIs are difficult to calculate, how curious genealogists can readily estimate indicative CIs for their own data, and why a growing number of genealogists are recognising that the inherent uncertainties which CIs quantify are so great that SNP-based TMRCAs are usually of much less practical use than is often assumed. The development of practical models that include inputs of STR and historical data can reduce these uncertainties, but the temptation to use and mis-interpret simplistic SNP-based TMRCA calculations is not going to disappear.

Introduction

The use of DNA data to calculate TMRCAs is a long-standing objective of the genetic genealogy community. The advent of Next Generation Sequencing (NGS) Y-DNA tests, such as FamilyTreeDNA's BigY test and the resulting SNP haplotrees, appear to offer a significant step towards this goal: many see SNP data as being more reliable than STR data, the calculation of SNP-based TMRCAs can be very simple, published SNP mutation rates are accompanied by a 95% confidence measure which, if not fully understood, seems to add comfort, and above all the resulting TMRCAs appear to be "in the right ball park". Though adopting very different approaches, both Dave Vance's SAPP model[2] and Iain McDonald's recent paper[3] appear to build on these lay perceptions and to supplement the basic SNP-based model with STR and historical inputs. However, the application of such models is too demanding for many administrators ("admins") of surname DNA projects, and as noted in his Conclusion, McDonald has not addressed some of the associated practical problems.

This paper addresses challenges presently facing project admins with limited mathematical skills and/or small samples when they attempt to estimate TMRCAs from SNP inputs derived from NGS tests such as BigY700. The following challenges are addressed in turn:

1. Understanding the basic maths, confidence intervals, accuracy and precision
2. Understanding average SNP mutation rates
3. Understanding SNP counts
4. Practical examples of SNP-based TMRCA calculations
5. Future developments

1. The basic maths

Calculating a predicted TMRCA can be very straightforward, even for those averse to mathematics. What we are exploring here is the use of SNP inputs (only)[4] to predict when the Most Recent Common Ancestor (MRCA) of two or more NGS testers was alive.[5]

---

[1] I am grateful to Robert Casey, Michael Cooley, Zack Doherty, Maurice Gleeson, David Hall, Jane Lindsay, Tom Little, Kathy McCauley, Dave Vance, Mary Wiley and Dennis Wright who kindly supplied data I used in Appendix B below. I am also very appreciative of comments on earlier drafts of this paper made by Maurice Gleeson, Iain McDonald, Ralph Taylor and Dave Vance, although my adoption of most of their valuable contributions does not imply they necessarily agree with my methodologies, opinions or conclusions.

[2] www.jdvsite.com, accessed 3 November 2021.

[3] McDonald 2021 at https://www.mdpi.com/2073-4425/12/6/862.

[4] I am only addressing SNP inputs to TMRCA calculations in order to keep the paper focussed and relatively short: introducing STR and historical data inputs, while clearly desirable, alas adds further complexity.

Conventionally TMRCAs are based on the years of birth of the tester(s) and of the MRCA. The former is usually assumed to be 1950CE = AD1950.[6] Thus for a single tester:[7]

TMRCA = (year of birth of tester) - (time t in years since the MRCA (a.k.a. coalescence age)

= AD(1950 - t), (1)

where t = r * n years, (2)

where r = relevant mean ("average") SNP mutation rate,
and n = count of SNPs since the MRCA.

For example, if r = 83 years per SNP and the tester has 5 Private SNPs since his Terminal SNP,[8]

t = 83 * 5 = 415 years; and in this case, where the ancestor characterised by this Terminal SNP,

TMRCA = AD1950 - 415 = AD1535.[9]

More generally, if several testers have descended from some MRCA,

t_{mean} = (mean SNP mutation rate) * (mean count of SNPs since MRCA) = $r_{mean} * n_{mean}$ (3)

But while this equation is mathematically correct, it gives a very deceptive impression of both accuracy and precision.[10] It is inaccurate because there are two basic methods of counting the number of SNPs since the MRCA, an issue addressed in section 3.3 below, and it is imprecise because SNP counts vary from patriline to patriline.[11] In fact both the components of equation (3) are Probability Distribution Functions (PDFs) of samples which contain inherent uncertainties. These uncertainties are conventionally represented by three basic equations, one for the estimated average or mean TMRCA,[12] which gives a single "point" date, and two which relate to the precision of this mean value and define its associated uncertainty: one for what is known as the Lower Confidence Interval (LCI) and one for the Upper Confidence Interval (UCI), thus

$t_{LCI} = r_{LCI} * n_{LCI}$ (4)

$t_{UCI} = r_{UCI} * n_{UCI}$[13] (5)

Confidence Intervals, like TMRCAs, are estimates, aka predictions. They are expressed in % terms, such as 95%, i.e. there is a 95% probability that the LCI and UCI calculated from a sample encompass the true TMRCA of the full population, and a 5% probability that they do not.[14]

Another important feature is that the product of two PDFs independent of one another is a PDF which will have CIs of 90% if these CIs are derived from the 95% CIs of the original PDFs,[15] thus:

$t_{90\%} \approx r_{95\%} * n_{95\%}$ (6)

And the product is a PDF with c.50% CIs if these are derived from the 68% CIs of the original PDFs:

$t_{50\%} \approx r_{68\%} * n_{68\%}$[16] (7)

---

5 This paper assumes a living person took the NGS test, as opposed to the testing of human remains ("ancient DNA").

6 CE = Common Era = AD; BCE = Before Common Era = BC. Ethnicity studies usually use "ybp" (years before present).

7 MRCA is defined by two or more testers, but the TMRCA of a single tester can be calculated if the MRCA SNP is known.

8 FTDNA effectively defines Private SNPs as those SNPs which have not yet been identified for any other tester, and the Terminal SNP as the most recent SNP currently shared by more than one tester.

9 YFull assumes testers are aged 60 (www.yfull.com < FAQ). McDonald (2021, 22) cited a sample poll which suggests testers had a mean age of 64 years and a 95% CI of 35-91 years. I gather this poll was taken in about 2014 and so has no impact on mean TMRCA estimates. Unlike the uncertainties in SNP mutation rates and SNP counts which have to be multiplied, these CIs are only additive, and as they are so small in the context of the other uncertainties, in practice they can be ignored. However the "drift" from dates such as these will become increasingly relevant as the years go by.

10 Accuracy is the closeness of observed data the true value; precision is the closeness of repeated data to each other.

11 SNP counts also vary considerably within each patriline (aka "lineage"), as shown by McDonald, the Clan Irwin Surname DNA project, and the MacAuley data used in Appendix B below.

12 For Normal PDFs (only), average = mean = mode = median.

13 Confusingly t_{LCI} gives $TMRCA_{UCI}$ and t_{UCI} gives $TMRCA_{LCI}$, but in practice this paradox does not affect the results.

14 A Confidence Interval, aka confidence limit, is twice the margin of error.

15 Conversely, if the 95% CIs of the product are required, these can be derived using the 97.5% CIs of the original PDFs.

It is important to recognise that equations (6) and (7) are both valid ways to quantify the same uncertainties: they simply express them in different forms. But which form is the more appropriate for TMRCA calculations in the genealogical context? The use of 95% CIs is customary in mathematics and some sciences,[17] and has been adopted by some genealogists. But as shown above, 95% CIs for SNP rates and for SNP counts give 90% CIs for TMRCAs, and there are several reasons why in practice 95% CIs for SNP counts are less predictable than has been appreciated hitherto, and why 68% CIs for SNP counts and the consequential 50% CIs for TMRCAs are more appropriate than 95/90% CIs:

- Appendix A below shows that few SNP counts have a clear "best fit" PDF, and the 68% CIs of the Normal, Poisson and Log-Normal PDFs differ from one another less than their respective 95% CIs, thus making the choice of PDF gives the "best fit" PDF as less critical;
- It also shows that unless the sample has a large number of testers, the 95% CIs derived from the actual cumulative frequencies of SNP counts are less reliable than similarly-derived 68% CIs;
- TMRCA 50% CIs span narrower date ranges than 90% CIs, and so are less likely to be irrelevant to conventional genealogical research (although of course they are also irrelevant 50% of the time);
- Fewer project admins should dismiss all CIs as academic niceties and/or remote contingencies.[18]

Another mathematical issue is the precision of the results of TMRCA calculations, and specifically the extent to which their results can or should be rounded. Given the value of TMRCA calculations currently perceived by many genealogists this is an important matter, although there seems to be no consensus. It is possible to calculate TMRCAs and their associated CIs to the nearest year[19] but as we will see, given the extensive number of uncertainties, several of which are not yet quantifiable, it seems appropriate to round off the results of calculations of TMRCAs to, say, the nearest 10 or even 50 years, and the associated lower and upper CIs to even more, perhaps to the nearest 100 years.[20]

While it is desirable to "round off" TMRCAs to avoid unjustified precision, this does not mean that the two components of SNP-based TMRCAs (SNP mutation rates and SNP counts) should only be calculated and cited to two significant figures, as this practice introduces avoidable and confusing errors, albeit that such errors may be trivial in the context of the underlying uncertainties.

2 Average SNP mutation rates

First, some more simple maths. Genetics theory tells us that

Average SNP mutation rate, r, in years per SNP = 1/(base pairs frequency * base pairs length) (8)

Several data sets are relevant:

Data set	base pairs' frequency	base pairs' length (hg38)	r, years per SNP
YFull (ComBED coverage)	$8.2*10^{-10}$ bps/year	8,482,579 bps	144.41
BigY500 (ex Warehouse, accessed 3 Nov. 2021)	($8.2*10^{-10}$ bps/year)	9,286,211 bps	131.32
FGC Elite 1 (ex Warehouse, accessed 3 Nov. 2021)	$8.2*10^{-10}$ bps/year	14,007,575 bps	87.06
BigY700 (ex Warehouse, accessed 3 Nov. 2021)	($8.2*10^{-10}$ bps/year)	14,626,759 bps	83.34
McDonald approximations (2021,3,23)	$8*10^{-10}$ bps/year	15,000,000 bps	"83" (83.33)

This table makes clear the important feature that there is no single "correct" average SNP mutation rate r: McDonald (2021, 3, 23-24) explains that the appropriate average rate depends on the length, measured in base pairs ("bps"), of the male-specific portion of the Y chromosome and on the

---

16 The product of two 68.3% CIs is a 47% CI, but for practical purposes I am assuming it to be 50% - an "evens" likelihood that the true mean is within these CIs and also, in this example, of it being outside these CIs. For Normal PDFs, 68.3% CIs = mean +/- SD, and 95.4% CIs = mean +/- 2*SD, and so the probability of a 95.4% CI is half that of a 68.3% CI.

17 For example 95% CIs are appropriate in ethnicity studies, because typical SNP counts and sample sizes are much larger.

18 These last two reasons are subjective: some may argue that the 50% CI ranges are still too wide to be of practical value to genealogists, and/or that a 50% chance that the CIs are irrelevant makes them valueless.

19 Until recently Alex Williamson's https://www.ytree.net was expressing TMRCAs to the nearest decimal of a year!

20 A parallel issue is how CIs are best expressed. Thus a TMRCA can be described as "AD1600, with 90% CIs of +200 years and -300 years" or "between AD1300 and AD1800 (90% CIs), with a mean of AD1600". Some even argue that it is preferable to omit the central year and only give a range of dates with their percentage CIs.

number of SNPs that are "callable". Thus for example if following a BigY test the relevant SNPs being counted are as analysed by YFull then the average rate of 144.4 years/SNP is appropriate, immaterial of the original test,[21] whereas if the SNPs being counted are those listed by FTDNA then the YDNA-Warehouse rates of 131.3 years/SNP are appropriate for BigY500 test results, and 83.3 years/SNP for BigY700 test results.[22] McDonald's rate of 83 years/SNP is an approximation, the absence of any decimal places no doubt being deliberately intended to reflect the underlying uncertainties.[23]

McDonald (2021, 24) lists a selection of published studies of modern average SNP mutation rates:

Paper	Reference in McDonald, 2021	base pairs' frequencies all in terms of 10^{-10} p.a. mean	95% LCI	95% UCI	skewed?	95% CIs as ratio of mean		68% CIs as ratio of mean	
Mendex et al.	[27]	6.17	4.39	7.07	left	0.288	0.146		
Adamov et al.	[5]	7.98	6.32	9.84	slightly right	0.208	0.233		
Poznik et al.	[30]	8.2	7.2	9.2	symmetric	0.122	0.122	0.06	0.06
Helgason et al.	[21]	8.33	7.57	9.17	slightly right	0.091	0.101		
Xue et al.	[28]	10	3	25	left	0.700	1.500		

From this he concludes that the callable SNP mutations in the Y-chromosome have a probable mean of c.8.2*10^{-10} base pairs p.a., and that the distribution about this mean is probably a Poisson PDF.

In section 2.2.2 of his paper McDonald considers different methods for handling the uncertainties of the SNP mutation rates for multiple tests. For his objective of a general model which also includes inputs of STR and historical data he treats the uncertainties in the mutation rates as he computes each SNP node in turn. For simplicity, and with less sophistication, this paper instead first explores the implications of recognising these uncertainties as a scaling factor over each cohort of SNPs. In this latter context the above data can be extrapolated thus:[24]

Analysis company/ databank	Test	Basis for Confidence Intervals	mutation rate r, years per SNP Mean	95% CIs Lower	95% CIs Upper	68% CIs Lower	68% CIs Upper
YFull	various	Adamov	144.4	120.6	178.1	128.5	160.3
FTDNA / Warehouse	BigY500	Poznik	131.3	115.3	147.3	123.4	139.3
FGC / Warehouse	Elite 1	Poznik	87.1	76.4	97.7	81.5	92.3
FTDNA / Warehouse	BigY700	Poznik	83.3	73.2	93.6	78.4	88.4

If the TMRCA being sought is to be based on SNP data for patrilineal descendants who have taken a variety of NGS tests, for example some BigY500 and some BigY700, then the SNP count for the Private SNPs in the BigY500 or BigY700 sample will need to have a correction factor applied.[25]

3 SNP counts

Four issues have to be considered when developing the relevant count of SNPs which are to be multiplied by the average mutation rates developed above:

1. Recognition that SNPs vary in quality and should be counted consistent with the test coverage.

[21] Note that YFull analyse FGC, FTDNA BigY500 and FTDNA BigY700 test data, but for all data they only count SNPs in the comBED regions before using their 144 years/SNP rate to calculate TMRCAs; in contrast the relevant rates published by YDNA Warehouse are applicable whenever the counts of SNPs called by FGC or FTDNA are used to calculate TMRCAs.

[22] See https://ydna-warehouse.org/coverage.html, accessed 3 November 2021. Note that the Warehouse data is updated from time to time and so the most up-to-date rates may differ slightly. For example, when accessed on 22 July 2021, BigY700 showed r = 83.38 years per SNP.

[23] For an alternative approach see www.jdvsite.com/faq < link to video on analysing BigY matches using SAPP.

[24] I have generated the mutation rates in this table by extrapolating the CIs in the previous table, which for the Poznik data curiously implies a symmetric PDF, not a Poisson PDF. McDonald used the Helgason data, which has slightly narrower CIs. Note also that the mutation rate relevant to an individual test depends on both the type of test and the exact coverage of the individual test, and I understand the latter can vary by about 10%. I am assuming, perhaps naively, that the 95% CIs in all these papers address the aggregate of all the relevant the uncertainties.

[25] Alternatively the data may be 'normalised' to some uniform bps length (e.g. as with the Royal Stewart data in Appendix B). Or the less reliable sample may be ignored (e.g. most would argue Y500 data is less reliable than Y700 data).

2. Recognition that the SNP counts currently available from NGS testing for each patriline are only a sample of those of the wider population of testers descending from their common ancestor;
3. Recognition of the bias inherent in most mean counts of SNPs since the MRCA
4. Recognition of the variety of methods of calculating the associated Confidence Intervals.

3.1 Recognition of the varying quality of Individual SNPs. While SNPs are not prone to the convergency/back mutation issues that bedevil STR data, and so can be considered much more stable, nevertheless what does or does not constitute a novel, callable, "phylogenetically significant" SNP is not clear-cut and unambiguous for every SNP. The quality of the callable SNPs to be counted depends on several factors, including:

- Coverage. The SNPs being counted must all occur in the same region of the Y chromosome as that where the chosen SNP mutation rate has been validated.
- Mapping. SNPs can be mis-called by the sequencer if two repeats are misaligned with each other
- Depth. For a SNP to be considered callable it must have a minimum number of reads, aka calls, typically 4, overlapping the location, but there is no fixed standard.
- Read consistency. Typically at least 90% of the reads should be derived rather than ancestral.

These features are beyond the scope of this paper, but the crucial point is that the criteria used for the SNP count should be the same as that used for the relevant test: if the criteria are different this may have a significant impact on TMRCA calculations.

FTDNA, YFull, Alex Williamson in his BigTree[26] and McDonald in his 2021 paper all consider the relevant VCF data or even BAM data for each SNP and make their own judgements as to what does or does not constitute a callable SNP.[27] There is an assumption that the FTDNA data stored in the YDNA Warehouse databank represents a consistent judgement on which SNPs are callable.

The project admin (or individual tester) has three options for handling this issue. They can:

- send the VCF or BAM data for the relevant test(s) to YFull (or to some private "expert"), bearing in mind that (i) although the Y-Full fee ($49) for a single NGS test is relatively trivial, the cost quickly increases as more project members subscribe, and (ii) the utility of this option is dependent on how many other matching NGS test results are already in relevant sections of the YFull haplotree; or
- analyse the VCF or BAM data themselves, as for example done by McDonald, or as explained by Vance.[28] This is a most satisfying and illuminating exercise, but difficult and laborious for the untrained, and involves project admins having to justify why their own analyses differ from those of YFull and/or FTDNA; or
- follow FTDNA's determination of which SNPs are callable for each individual BigY test in light of their ever-increasing awareness of the haplotree of mankind.[29] FTDNA frequently refine these details, which means that the haplotree and Private SNPs that appear on their web pages are kept updated as the haplotree matures, but are liable to change from time to time.

The vast majority of testers and project admins choose the third of these options, albeit perhaps unconsciously.

But whichever option is chosen, there is a need to periodically review the SNP count that has been used to calculate a TMRCA.

3.2 Recognition that SNP counts are samples of a larger population. While it is tempting to regard SNP counts as deterministic, and to assume the TMRCA back to some designated SNP that can be calculated without any probability caveats such as CIs, most TMRCA calculations involve more than one tester sharing descent from some MRCA SNP, and the SNP count frequencies are inherently some form of PDF, even if it is not necessarily a close fit with one of the more common continuous

[26] See https://www.ytree.net. Dennis Wright has similarly processed the O'Brian and R-L226 data in Appendix B below.
[27] The ISOGG Y haplotree has yet another set of criteria for what constitutes a phylogenetically-significant SNP.
[28] See www.jdvsite.com/faq < link to video on analysing BigY matches using SAPP.
[29] This is additional to their routine "naming" of previously un-named SNPs that were "Private" to another tester.

PDFs. Nor are such SNP counts static, for they evolve over time as more such descendants take the relevant NGS test, and as the "callability" of marginal SNPs evolves in response to improved understandings of SNP quality, as described above). In fact the available relevant SNP counts of most patrilines are samples of larger populations of descendants, typically of unknown size, and this introduces inherent uncertainties in the available evidence of SNP counts.

3.3 Recognition of the bias inherent in most mean counts of SNPs. The use of the mean count of SNPs since the MRCA when calculating TMRCAs, as in equation (3) above, is a convenient and popular method.[30] In contrast YFull and McDonald[31] apply a more refined method of counting SNPs since the MRCA, using an unbiased node-by-node SNP count. This method involves a sequential, bottom-up counting process for each node of the haplotree in turn. The methodology is best understood by way of an example – see section 4.1 below. If the relevant haplotree is fully symmetrical in shape then the two methods will give the same resultant count of SNPs since the MRCA, but of course all but the very simplest of haplotrees are asymmetrical and so in practice the two methods give different SNP counts.[32] Several issues thus arise:

- Why do the two methods give different results? This is because the mean SNP count method introduces a bias at each node in the haplotree if the patrilines below this node represent different numbers of testers; for example, if there are three patrilines below a node, of which one represents three testers, one represents two testers and the other represents a single tester, then by the mean count method the first two patrilines give undue weight to the SNP counts of their respective testers. Again this is best understood by the example in section 4.1 below. It follows that this node-by-node method avoids the biases that are inherent in averaging such data, and is thus clearly more accurate than the more convenient mean SNP count method.
- Why then is use of this unbiased node-by-node method not more popular? The biases that this method avoids have only been publicised relatively recently and so few project admins are aware of it, and of those who are, some do not recognise its greater accuracy. It is also more complicated and laborious to apply, and, especially for large sample sizes or if the haplotree is not redrawn to show the nodes clearly, is more prone to errors during application.[33] It can also be argued that the difference in the results of the two methods is likely to be well within the associated Confidence Intervals,[34] and so the extra effort is not justified. Nor, because of its relative complexity, is this node-by-node likely to become more popular than the mean count method in the future, unless its complexity can be circumvented by some user-friendly software.
- How do the results of the two methods differ? Because the difference for each sample will depend on the shape of each relevant haplotree it is not possible to predict the size of the difference, or even to develop some simple "rule of thumb" to forecast which method will give the larger SNP count, as the following examples show:[35]

Sample			Sample size	SNP count		Difference		
			(no. of testers)	convenient mean	node-by-node	SNPs	%	TMRCA
Royal Stewart	S781	var	26	5.96	6.86	-0.90	-13%	-113 years
Border Irwin	FGC13746	Y700	65	7.57	6.72	+0.85	+13%	+71 years
Lae/Lay	FT21692	Y700	34	3.56	2.79	+0.77	+28%	+64 years
Doherty	BY472	Y500	30	10.38	10.83	-0.45	- 4%	-60 years
MacAuley	Y179697	Y700	32	7.59	6.88	+0.71	+10%	+59 years
Doherty	BY471	Y700	63	12.90	13.60	-0.70	- 5%	-58 years
Irwin	FT104360	Y700	6	13.33	12.83	-0.50	- 4%	-42 years

[30] For examples of this method see Holton GD *Tracing your Ancestors using DNA*, Barnsley 2019, pp.137-9, and Dave Vance's SAPP model; the latter uses a hybrid method: although SNPs are counted at each node, the mathematical result is the mean SNP count, to which is applied an arbitrary weighting to address outliers (see www.jdvsite.com/faq < link to video analysing BigY matches using SAPP, accessed 3 November 2021). Note both sources caution that the results of their TMRCA calculations should include a margin of error of a couple of centuries either way.

[31] www.yfull.org; McDonald 2021.

[32] The two methods also give the same result if the TMRCA is between a single tester and one of his ancestors.

[33] This will remain true until the process can be encapsulated into a sophisticated computer program.

[34] This statement is correct, although the differences relate to accuracy whereas the CIs relate to precision.

[35] This data was collected before I had appreciated the significance of biases attributable to the haplotree shape.

More study is needed to clarify this issue, but meanwhile we can note that these differences imply that mean count TMRCA calculations may incorporate errors of +/- c.1-3 generations.

3.4 Recognition of the variety of methods of calculating associated Confidence Intervals. Although it has long been recognised that calculations of predicted TMRCAs should always be accompanied by their associated CIs, and that such CIs can be characterised, at least conceptually, by some common type of PDF, in practice such CIs are very rarely calculated. There are several reasons for this: few individual testers and project admins are interested; some argue that they know the range of CIs is so large that it would be a waste of time to calculate them; a small minority who are curious find the methodology unclear and the maths too complicated; and even the most diligent analysts struggle to access samples of SNP counts that are usually not publicly available. And, I now appreciate, it was naïve to expect that SNP counts can be reliably represented by one of the common PDF types. So hitherto little attempt has been made to use empirical data to determine how SNP-based TMRCA CIs should be derived. This subject is developed further in Appendix A below, whose findings in this context may be summarised thus:

- There is no single, "text book" method with which SNP-based TMRCA CIs should be calculated.
- Even large SNP counts cannot be represented by any common PDF because of "noise" attributable to features such as asymmetrical haplotrees, population size, family size, father's age etc.
- Of the more common PDFs, the Poisson PDF is the "best fit" (though not necessarily a "good fit") to the samples of SNP counts within the surname era[36] that have been analysed.
- The CIs associated with any "best fit" PDF are not necessarily the most reliable guide to the CIs associated with such counts. CIs can also be interpolated from the actual cumulative frequencies of the SNP counts in the sample, except when the number of testers is small, these CIs can offer a more reliable method than the CIs derived from "best fit" PDFs.
- So conceptually the choice of the appropriate CIs for each specific sample of SNP counts could be determined by identifying the PDF giving the best "best fit" to this data, then comparing the CIs derived from this "best fit" PDF with the CIs derived from the actual cumulative frequencies of the SNP counts, and finally making a subjective choice of the most appropriate CIs.

Although pragmatic, such a conceptual process has many disadvantages: it is laborious and error-prone, it is impractical for TMRCA calculations when fewer than c.15 testers share a MRCA SNP, it is unattractive to genealogists lacking the necessary patience or understanding of statistical theory, and after all the effort the chosen CIs are not necessarily as objective or accurate as the calculations imply. To avoid all but the last of these issues I have developed two simple, empirical formulae which can be readily used by non-mathematically minded genealogists for all SNP-based TMRCA calculations within the surname era, even when SNP counts of only a few testers are available:

$$\text{Estimated Lower } CI_{68\%} \text{ of SNP count} \approx \text{mean SNP count} - (\text{sq.root mean SNP count} * TF_L) \quad (8)$$

$$\text{Estimated Upper } CI_{68\%} \text{ of SNP count} \approx \text{mean SNP count} + (\text{sq.root mean SNP count} * TF_U) \quad (9)$$

where TF_L and TF_U are factors derived from the following table for 68.3% CIs:[37]

No. of testers (N):	2	3	4	5	6	7	8	9	10	12	15	20	30	50	100	∞
$TF_{L\,68.3\%}$:	1.85	1.32	1.20	1.14	1.11	1.09	1.08	1.07	1.06	1.05	1.04	1.03	1.02	1.01	1.01	1.00
$TF_{U\,68.3\%}$:	2.775	1.98	1.80	1.71	1.67	1.64	1.62	1.61	1.59	1.58	1.56	1.55	1.53	1.52	1.52	1.50

These formulae are used to estimate the CIs on the bottom line of the entries for each sample in column O of the Summary table in Appendix B. The resulting CI estimates can be seen to be pretty close to the CIs derived from the "best fit" CIs and those interpolated from the cumulative frequencies; the few that are narrower are shown in *italic* font. This implies that however crude and

[36] By surname era I mean the period since when surnames first became hereditary, typically c.600-1,000 years ago.

[37] In this table TF_L is copied from the similar table introduced in Appendix A below. TF_U is simply (TF_L * 1.5), where 1.5 is an arbitrary, empirical factor to allow for the longer right-tail of the Poisson PDFs which Appendix A has shown to be characteristic of most SNP counts within the surname era. Similar simple formulae could readily be developed for 95% CIs, but the simplifications would introduce appreciable errors and also give a misleading impression of accuracy.

illogical these simple formulae may be, they nevertheless offer a fair "rule-of-thumb" indication of the uncertainties associated with SNP counts, at least until a more refined substitute is developed.

So subject to various assumptions already addressed, equations (3) - (6), (8) and (9) above can be modified to give three simple, ubiquitous equations, thus:

$$TMRCA_{mean} = AD(1950 - (r * n)) \quad (10)$$

$$TMRCA_{50\%LCI} \approx AD(1950 - (r_{68\%UCI} * (n + (sq.root\ n * TF_U)) \quad (11)$$

$$TMRCA_{50\%UCI} \approx AD(1950 - (r_{68\%LCI} * (n - (sq.root\ n * TF_L)) \quad (12)$$

Clearly equations (11) and (12) only offer approximate estimations and are no substitute for mathematical rigour.[38] The three equations can be represented by the following "look-up" table:[39]

Table for estimating TMRCAs based on mean SNP counts and associated indicative 50% Confidence Intervals for BigY700 testers

Year AD of TMRCA & 50% CIs		No. of testers (N) descended from MRCA											
		2		4		6		10		30		∞	
Mean SNP count since MRCA (n)	Predicted TMRCA	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI	Lower 50% CI	Upper 50% CI
1.0	1867	1616	2017	1702	1966	1714	1959	1721	1955	1726	1952	1729	1950
2.0	1783	1426	1998	1548	1926	1565	1916	1574	1911	1582	1906	1586	1904
3.0	1700	1260	1966	1409	1878	1430	1866	1441	1859	1451	1853	1455	1851
4.0	1617	1106	1926	1278	1825	1302	1810	1315	1803	1326	1796	1331	1793
5.0	1534	959	1882	1152	1768	1179	1753	1194	1744	1206	1737	1211	1733
6.0	1450	819	1835	1030	1710	1059	1693	1075	1683	1088	1675	1095	1672
7.0	1367	682	1785	910	1650	942	1631	959	1621	973	1613	980	1609
8.0	1284	549	1733	793	1589	826	1569	845	1558	860	1549	868	1545
9.0	1200	418	1680	677	1527	713	1505	733	1494	749	1484	757	1480
10.0	1117	290	1625	563	1464	601	1441	622	1429	638	1419	647	1414
11.0	1034	164	1569	450	1400	489	1376	511	1363	529	1353	538	1348
12.0	950	39	1512	338	1335	379	1311	402	1297	421	1286	430	1281
13.0	867	BC84	1454	227	1270	270	1245	294	1230	313	1219	323	1213

NB Above dates are given to nearest year to avoid interpolation errors;
after interpolating for intermediate values of mean SNP count and no. of testers,
dates of TMRCAs should be rounded to nearest 10 years, and
dates of 50% Confidence Intervals should be rounded to nearest 100 years

Assumptions: Testers born AD1950; mean SNP mutation rate: 83.3 yrs per SNP, 68% CIs 78.4 - 88.4 yrs/SNP;
Lower 68% CI of mean SNP count n = sq.root n*(t factor); Upper 68% CI = 1.5*Lower CI.

This table enables curious genealogists to derive indicative CIs associated with TMRCAs derived from mean SNP counts, and to keep such TMRCAs in perspective. But more importantly it demonstrates clearly that even at best, 50% of the time the mean TMRCA will be outside a CI range of at least two centuries, and typically of much longer periods. And of course 90% CIs cover even wider periods.[40] This confirms the unpalatable and widely unrecognised fact that, pending possible future developments (see section 5 below), TMRCAs derived from SNPs alone are of much more limited practical use to genealogists than many of them assume.

Strictly speaking equations (11) and (12) and the above table are not applicable if TMRCAs are calculated with the node-by-node method to avoid biases, although it could be argued that the difference between mean-based and unbiased node-by-node based SNP counts is a matter of accuracy whereas the associated CIs are a matter of precision. On the other hand the calculation of CIs for node-by-node based TMRCAs would be tortuous without a software package.

[38] For example, in theory CIs scale as sq.root n/(N-1), but this is offset by the SNP counts not being fully independent.

[39] The shaded dates are of Upper 50% CIs which are later than the mean birth dates of the testers. Similar tables could be developed for BigY500. Predicted TMRCA accuracy is improved by the node-by-node method (see section 3.3 above).

[40] The same can be said for TMRCAs based on STRs alone.

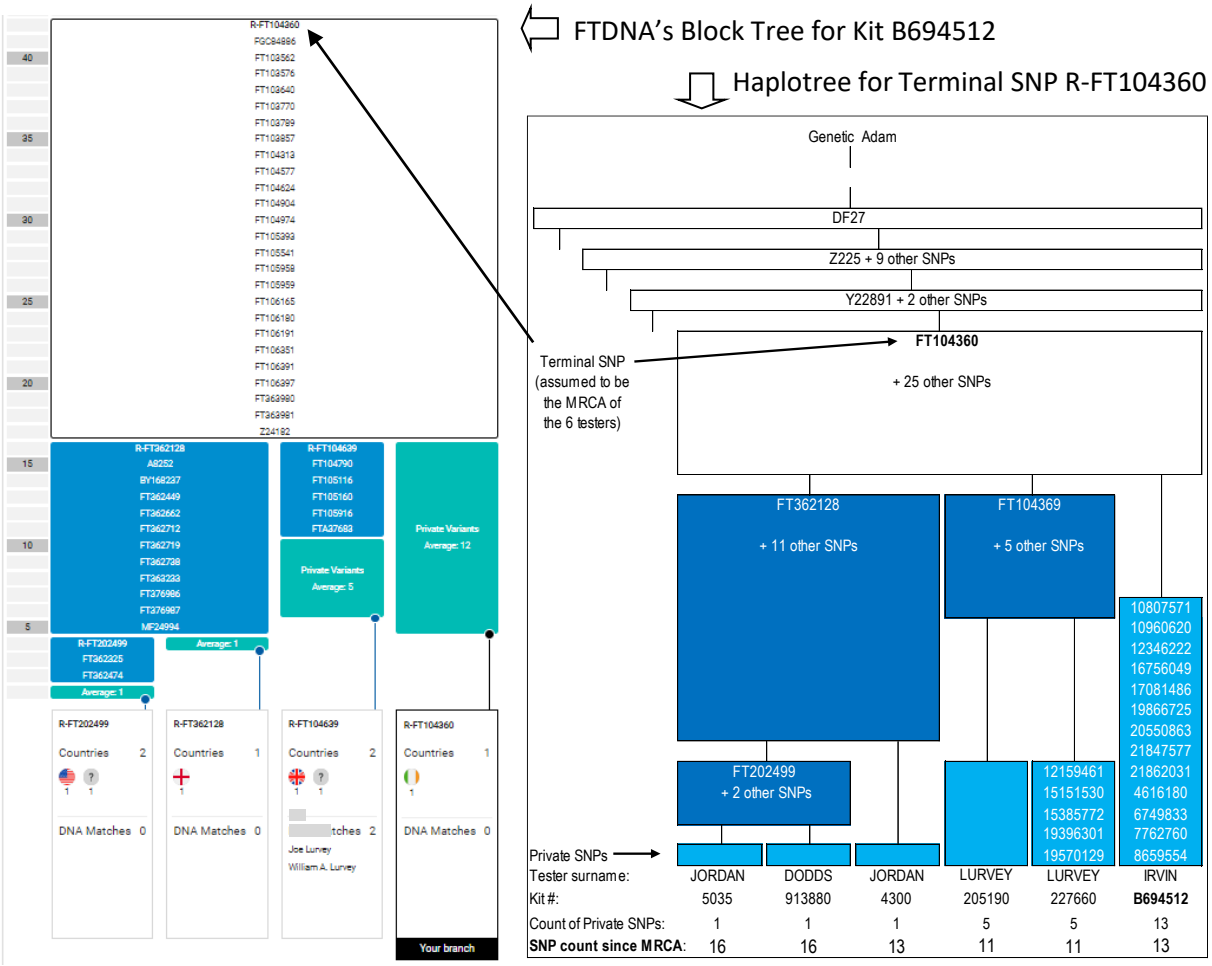
4 Practical examples of SNP-based TMRCA calculations

The following examples use data from the Clan Irwin Surname DNA project[41] to illustrate some practical applications of the above considerations when calculating TMRCAs from SNP data alone.

4.1 Predicting TMRCAs from data in FTDNA's Block Trees

This is a common challenge for FTDNA's project admins. Below is a copy of the Block Tree for a Clan Irwin tester kit B694512 who has only one YDNA "match" in the Irwin Surname DNA project, who in turn happens to be a tester with the surname Lurvey. The question arises of whether one of these two "matches" has an NPE[42] in his ancestry, or whether their shared Terminal SNP, identified by FTDNA as R-FT104360, is so old that it probably represents a common ancestor who lived before the surname era.

The first step is to aggregate as much relevant SNP data as possible[43] from publicly available sources to build the haplotree downstream of the Terminal SNP R-FT104360.[44] By referring to FTDNA's public web pages for the Jordan, Dodds and Lurvey surname projects, the Block Tree for kit B694512 (below, left) can be expanded into a haplotree for SNP R-FT104360 (below, right):



All six of these men have inherited all 26 SNPs in the R-FT104360 block, although pending more BigY test results we do not know which of the SNPs in this block is the most recent. Nevertheless if we count the SNPs in each patriline subsequent to this block we can calculate the likely TMRCA of the 6 men thus:

41 See www.clanirwin-dna.org.

42 For a discussion of NPEs see www.isogg/org.wiki < Non-paternity event.

43 The more SNPs that can be included in the TMRCA calculation the less unreliable the calculation will be.

44 FTDNA's Block trees are limited to 30 matches, and so project admins may have to refer to more than one Block tree to build the relevant haplotree.

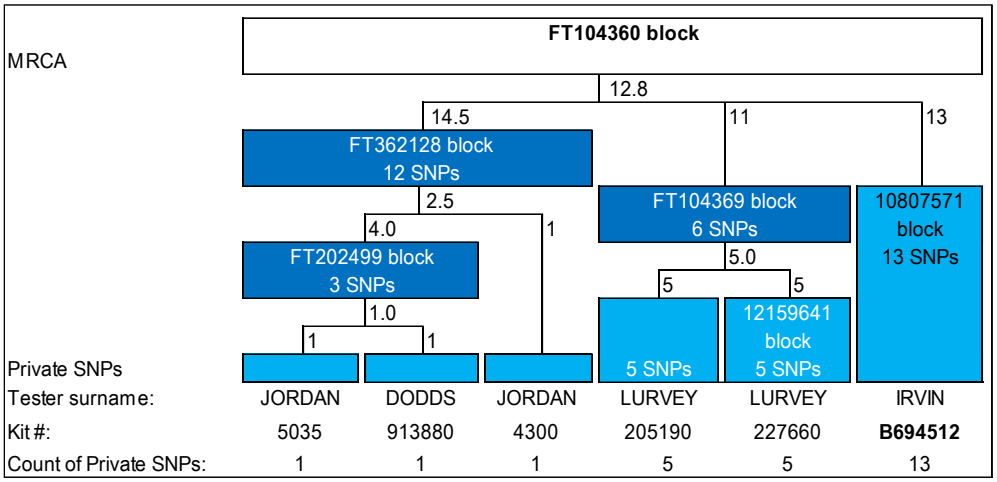
total SNP count = ∑n = (16 + 16 + 13 + 11 + 11 + 13) = 80 SNPs[45]

mean SNP count = (∑n)/n = 80/6 = 13.3 SNPs

age to coalescence = t = r * n_{mean} = 83.3 * 13.3 = 1111,[46] say 1110 years

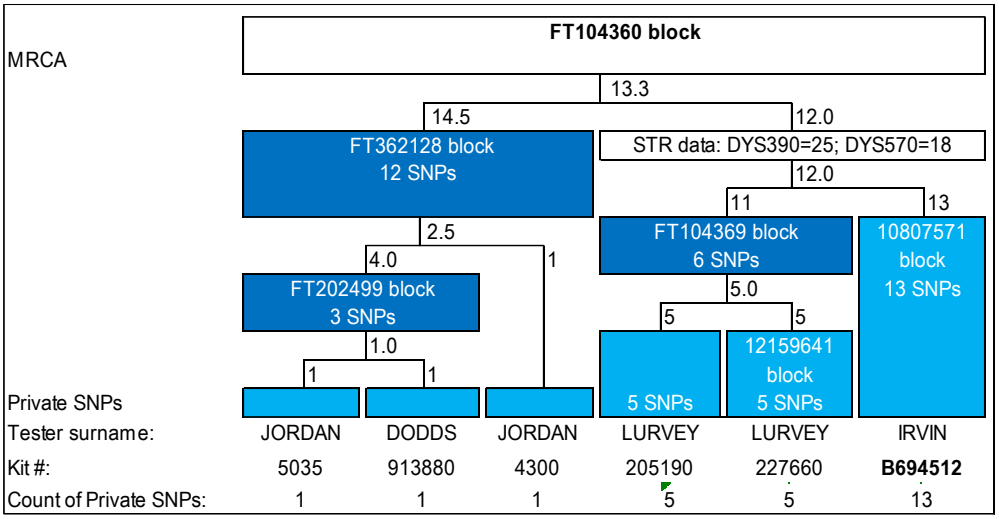
and hence TMRCA ≈ AD1950 - 1110 = AD840.

However this calculation is biased by the dominance of the 3 men sharing descent from FT362128. Such biases are avoided by instead adopting the more refined node-by-node method addressed in section 3.3 above. This method can be applied to this example by adapting the haplotree thus:



The pros and cons of node-by-node calculations versus the more convenient averaging of SNP counts have been developed in section 3.3 above. The implications for this particular example are that the MRCA SNP count changes from 13.3 to 12.8 SNPs, and thus the TMRCA from AD840 to AD880.

A further refinement, albeit outside the nominal scope of this paper, is to adapt this haplotree into a mutation history tree[47] by adding some STR data. Inspection of the FTDNA public pages for these 6 men shows that the Lurveys and Irvin share different counts for two STRs, DYS390 and DYS570, from the other 3 men. This may be represented thus:



The MRCA count thus changes again, by coincidence from 12.8 back to 13.3, and the resulting TMRCAs from AD880 back to AD840. But whether the predicted TMRCA is AD840 or AD880, both these dates predate the surname era, and so it is unlikely that either the Irwin or Lurvey patrilines

[45] Why the "Block Tree" for B694512 shows 12 Private SNPs but his "Results" show 13 is unclear. I note similar minor discrepancies from time to time, and regard FTDNA's "Results" data as being more reliable than their "Block Tree" data.

[46] We know that all six testers took the BigY700 test because they share a terminal SNP prefixed by "FT", which is effectively specific to SNPs identified by BigY700 tests.

[47] A mutation history tree is a haplotree extended to include relevant STR data.

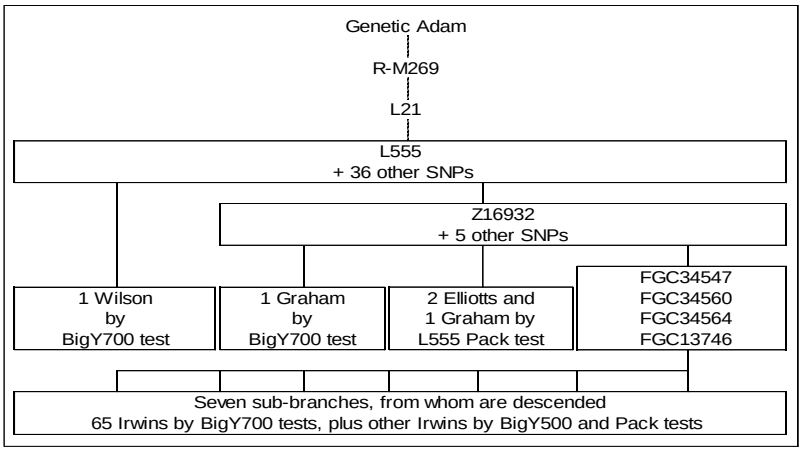
included a NPE that adopted the other's surname.[48] Instead it is more likely that each patriline descending from the R-FT104360 block included one or more SNP mutations which predated the surname era.

4.2 Can we identify the SNP characterising the MRCA within a SNP "block".

The small branch of Irwins descended from R-FT104360 is one of over 40 branches of this surname identified by the project. The largest of these branches is the Borders branch, which currently has 65 BigY700 testers.[49] Some descendants of this branch still live today near the Scottish Borders, and a few have patrilines dating back to ancestors who lived in Dumfriesshire c.1500. Alas we know neither the name nor dates of the founder of this branch, but by combining some historical evidence with the R-L555 haplotree we can identify the SNP "block" which is most likely to include the SNP representing the earliest Border Irwin. The immediate challenge is to see if we can tentatively identify which SNP within that block is most likely to be representative of the MRCA.

Let us first consider the historical evidence. The surname de Irwyn occurs across central Scotland in the 13th century, but the lengthy Ragman Roll of 1296 (in which no Irwins appeared) suggests that the use of hereditary surnames across Scotland was then still only common amongst the nobility. The earliest extant use of de Irwyn as a hereditary surname was in Aberdeenshire during the 14th century. When extant records become prolific in the Scottish Borders in c.1500 we find evidence of several contemporary branches of the surname in Dumfriesshire which appear to be loosely related to each other, suggestive of a common ancestor a few generations earlier. There is some more specific evidence: a Nicholas de Irwin who was briefly a vicar at Buittle in the 1370s, and John and Gilchrist de Irwin who were tenants at Buittle and Morton respectively in 1376.[50] It is possible that Nicholas was the father of John and Gilchrist, and if so that Nicholas was born c.1320, his sons were born c.1350, and it would have been his grandsons born c.1380 who were the first Irwins to be born in Dumfriesshire. But this is speculation, for we have no evidence that Nicholas, John or Gilchrist had surviving offspring, or that one or more were an ancestor of later Border Irwins. The challenge is thus to see if ySNP data can throw any light on these uncertainties.

When the SNP R-L555 was discovered seven years ago by a Walk-the-Y DNA test it seemed possible that this SNP might be unique to the Border Irwins. It is now apparent that this SNP is shared by some Wilsons, Grahams and Elliotts, which like Irwin are surnames common on the western Border, and that it is a block of four SNPs younger than R-L555 which are unique to the Border Irwins:[51]



Pending further NGS tests which may break down this block of four SNPs, we do not know the sequence in which their respective mutations occurred, and so the sequence in which these SNPs

[48] This is (just) so even after considering the Upper 50% CI, in this case c.AD1300 (from table in section 3.4 above).

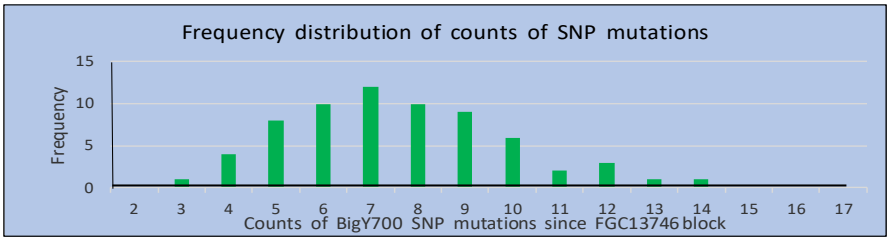
[49] 9 further Border Irwins have taken BigY500, but their inclusion in this section would add unnecessary complication.

[50] Buittle and Morton are 10-15 miles west of Dumfries. Although a Gilchrist filio Eruni was a witness in Dumfries in 1124x1185 and two Irwins lived in Berwick on Tweed c.1330, the former cannot have been a hereditary name, and there is no evidence that the latter had descendants. For more details see Irvine *The Irwin Surname* 2020, 65-9, 139.

[51] "Border Irwins" include various spellings of the name sharing descent from R-FGC13746, and NPEs sharing this SNP.

are listed, whether by FTDNA or anyone else, is quite arbitrary, as is the SNP chosen to identify the block. Nevertheless the TMRCA of descendants of this block can be readily calculated, as can the date that this block was formed.

The following bar chart shows the current count of BigY700 SNPs downstream of R-FGC13746 block:

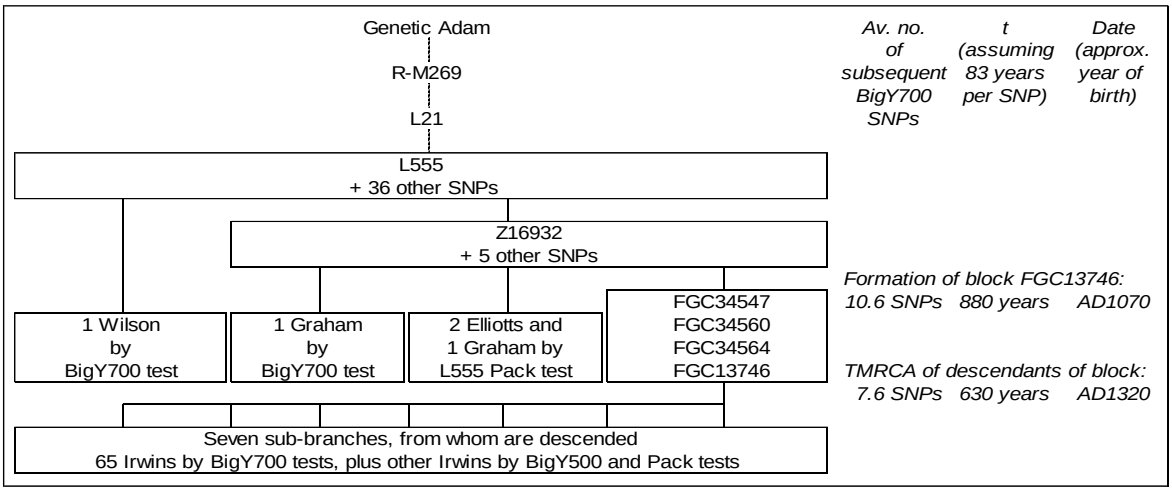


The mean count of SNPs since the FGC13746 block is 7.6,[52] and so

$t_{mean} = n * \mu = 83.3 * 7.6 = 633$ years, say 630, and hence

TMRCA ≈ AD1950 - 630 = AD1320.

Similarly the mean count of SNPs since the earliest SNP in this block is 10.6 SNPs, giving date of the formation of this block as AD1070.[53] We thus have, prima facie:



So if the date range of the FGC13746 block is AD1070 to AD1320, it seems the youngest SNP in this block might represent the earliest Border Irwin, even if we don't know its identity. However the more accurate node-by-node method gives the range of SNP counts for the FGC13746 block as 6.7 to 9.7, and hence the date range as AD1140 to AD1390, suggesting it might be the second youngest SNP in this block that best represents the earliest Border Irwin, even if its identity is also unknown.

It is thus apparent that "No", we cannot determine exactly which SNP within this block characterises the earliest Border Irwin. And while it may seem comforting that the above TMRCAs are at least compatible with the historical evidence,[54] it is important that even this tenuous impression needs to be placed in the context of the associated confidence intervals.

---

52 Calculated in Appendix C.

53 YFull apply the term "formed" to the describe age of the oldest SNP in a block, in this example 880 ybp, or AD1070.

54 It is curious that this Border Irwin example of a SNP-based TMRCA matching closely to historical data is not unique:

Sample		TMRCA based on			
		Historical data only	convenient mean SNP count	node-by-node SNP count	McDonald 2021, 19
Royal Stewart	S781	c.1245	c.1200	c.1090	c.1254
Border Irwin	FGC13746	c.1350	c.1320	c.1390	-
MacAuley	FT22697	c.1766	c.1760	c.1820	-

No doubt research into other surname branches can reveal similar examples. But we have to recognise that such examples are fortuitous coincidences rather than conclusive evidence. It is also coincidental, and ironic, that in these three examples the convenient mean-based TMRCAs appear more accurate than the node-by-node based TMRCAs!

4.3 Estimating approximate Confidence Intervals ("CIs")

Section 3.4 above has shown these calculations can be made quite simple. For example, using the Border Irwin R-FGC13746 data introduced above and equations (10), (11) and (12) above we have:

$TMRCA_{mean}$ = AD(1950 - (r_{mean} * n)), where r_{mean} = 83.3 yrs/SNP and n = 7.57 SNPs
= AD(1950 - (83.3 * 7.57)
= AD(1950 - 631)
≈ AD1320[55]

$TMRCA_{50\%LCI}$ ≈ AD(1950 - ($r_{68\%UCI}$ * (n + (sq.root n * TF_U)), where $r_{68\%LCU}$ = 88.4 yrs/SNP & TF_U = 1.52
≈ AD(1950 - (88.4 * (7.57 + (2.75 * 1.52))
≈ AD(1950 - 1034)
≈ AD920

$TMRCA_{50\%UCI}$ ≈ AD(1950 - ($r_{68\%LCI}$ * (n - (sq.root n * TF_L)), where $r_{68\%LCI}$ = 78.4 yrs/SNP (ex section 2), sq.root n = 2.75, and N = 65, so TF_L = 1.01 (ex section 3.4)
≈ AD(1950 - (78.4 * (7.57 - (2.75 * 1.01))
≈ AD(1950 - 376)
≈ AD1570

i.e. TMRCA = AD1320 (50% CI: 920-1570).

Similar dates can be interpolated from the table at the end of section 3.4 above.

For those who are averse to using approximate tools such as these, Appendix C shows that the "best fit" PDF for this data is a Poisson PDF, and entering 7.57 as the number of "Observed Events" at www.statology.org/poisson-confidence-interval-calculator gives 68.3% CIs of 4.9 and 11.5, so we have:

$TMRCA_{50\%LCI}$ = AD(1950 - ($r_{68\%LCI}$ * $n_{68\%LCI}$)
= AD(1950 - (78.4 * 4.9)
= AD(1950 - 384)
≈ AD1570

and $TMRCA_{50\%UCI}$ = AD(1950 - ($r_{68\%UCI}$ * $n_{68\%\ UCI}$)
= AD(1950 - (88.4 * 11.5)
= AD(1950 - 1016)
≈ AD930

i.e. TMRCA = AD1320 (50% CI: 930-1570)

and Appendix C also shows that the actual cumulative frequencies give 68% CIs of 4.6 and 10.4, so:

$TMRCA_{50\%LCI}$ = AD(1950 - ($r_{68\%LCI}$ * $n_{68\%LCI}$)
= AD(1950 - (78.4 * 4.6)
= AD(1950 - 361)
≈ AD1590

and $TMRCA_{50\%UCI}$ = AD(1950 - ($r_{68\%UCI}$ * $n_{68\%\ UCI}$)
= AD(1950 - (88.4 * 10.4)
= AD(1950 - 919)
≈ AD1030

i.e. TMRCA = AD1320 (50% CI: 1030-1590).

This latter method for calculating TMRCA CIs is probably the most reliable for this sample, and for most samples with mean SNP counts of > c.15, but the first method is the easiest to calculate, and is also the only method applicable to samples with mean SNP counts of < c.15. But even with all this attention we almost certainly have not yet captured all the subtleties of SNP mutations, so it is probably safest just to say that for this sample:

TMRCA = AD1390 (50% CI: c.900-1600).

While CIs such as these "flag" the uncertainties inherent in all SNP-based TMRCA models, this example exemplifies the warning in section 3.4 above that even 50% CIs render TMRCAs based on SNPs alone to be of very limited practical use to the genealogist. It is thus understandable that some project admins advise that all SNP-based TMRCA calculations should be taken with a large "pinch of salt", or even entirely ignored.

[55] As we have seen, the more accurate node-by-node method gives a TMRCA of AD1390 for this data.

5 Future developments

This paper is not suggesting that this disappointing demonstration of the severe limitations of SNP-based TMRCAs is the last word on this subject, or that the empirical equations (11) and (12) for estimating TMRCA CIs are a substitute for mathematical rigour. Several future developments can be identified that should help to narrow the CI ranges that measure the uncertainties inherent in all TMRCA calculations:

- Both Vance and McDonald have already demonstrated the ability to combine SNP inputs with STR and historical data inputs for calculating TMRCAs for genealogists. Both these models represent substantial advances, even if at present the former does not avoid the biases inherent in asymmetrically shaped haplotrees and lacks quantified CIs, while the latter is still accompanied by practical problems, and its inherent complexity makes the associated maths too advanced for use by most genealogists. Significant refinements in both models are thus clearly feasible, albeit the challenges remain formidable.
- FTDNA are expected to soon improve their on-line resource for deriving TMRCAs and thus hopefully contribute a significant step forward for most project admins.
- The pending YDNA Warehouse website platform will hopefully enable much more extensive and rigorous analyses of empirical SNP data than has been possible in Appendix A to this paper, thus leading to improvement in the understanding of the characteristics of various features that may influence SNP mutations.
- New DNA testing technologies and analysis tools may also help refine TMRCA calculations.

Notwithstanding these likely developments, SNP data will continue to contribute to TMRCA calculations, the need for accompanying CIs will not cease, and many of the principles addressed in this paper will remain relevant. And nor is the simplistic attraction of multiplying SNP counts by some mean SNP mutation rate to derive misleading TMRCAs likely to disappear.

6 Conclusions

This paper addresses several important conclusions associated with SNP-derived TMRCAs:

1. The mathematically simple equation of TMRCA = AD1950 - (r * n), where r is an appropriate average SNP mutation rate and n is some relevant SNP count, gives a very deceptive impression of both accuracy and precision, unless accompanied by appropriate confidence intervals (CIs).
2. The appropriate r to use, such as 144, 131 or 83 years per SNP, depends on the circumstances.
3. There are two basic methods of counting SNPs when calculating TMRCAs; these may be described as the mean SNP count method and the unbiased node-by-node method. The former is better known and more convenient to use; the latter is more accurate as it avoids biases due to haplotree asymmetry. TMRCAs calculated by the two methods may differ by 1-2 generations.
4. TMRCAs are liable to change over time as more descendants test and as callable SNPs are refined.
5. SNP rates and counts are samples of larger populations and are thus probability distribution functions (PDFs). The more testers whose SNPs can be counted, the more precise their calculated TMRCAs become.
6. TMRCAs calculated using SNP rates and SNP counts with 95% CIs have 90% CIs, and those calculated using 68% CIs have c.50% CIs.
7. For several reasons it is arguable that TMRCAs with 50% CIs (derived from SNP rates and SNP counts with 68% CIs) give genealogists a more appropriate guide than TMRCAs with 90% CIs.
8. TMRCAs should be rounded to the nearest 10 or 50 years and their associated CIs to 100 years.
9. There is no clear “correct” method for calculating CIs of SNP counts. Most SNP counts within the surname-era have a Poisson PDF “best fit”, but in many cases this “best fit” PDF gives a narrower range of CIs than those derived from the cumulative frequencies of actual SNP counts, when these can be calculated, i.e. when the number of testers exceeds about 15.

10. Equations (11) & (12) and the table in section 3.4 above offer tools for deriving indicative CIs that are simple for curious genealogists to apply, less prone to calculation errors, and, uniquely, are applicable when the number of testers is small. The validity of these tools is confirmed by their application in Appendix B below.

11. Notwithstanding the above caveats and refinements, the range of CI values associated with SNP-based TMRCA models is often so great that alone these models are of much less practical use than many genealogists believe.

12. Appendix A below also shows that the expectation that the probability distributions of the larger samples would tend towards a smooth bar chart and a close fit to one of the common PDFs that might be expected from the random mutations of SNPs was naive. Instead the "spikiness" of these SNP counts is indicative of "noise" caused by other factors such as biases inherent in the asymmetric shape of most haplotrees, changes in population size, varying family sizes, the age of the father, Founder effect etc.

13. Several potential developments may help alleviate these otherwise depressing findings. These developments could include further refinement of Vance's SAPP tool, a more practical version of the model recently developed by McDonald, and the pending improvements in the current FTDNA and YDNA Warehouse websites. How much such developments may improve the accuracy and precision of TMRCA calculations remains to be seen, but they will not remove all the underlying uncertainties and nor will they override the underlying principles addressed in this paper, or make the superficial attraction of simplistically derived SNP-based TMRCAs disappear.

14. Pending such developments, all TMRCAs, not least those based on SNP inputs alone, even when accompanied by CIs, remain a very crude tool for the typical genealogist, unless confirmed by a reliable independent source.

Appendix A - Statistical analysis of SNP Counts

To date most genetic genealogists have given little consideration to the various statistical characteristics of the SNP counts that many surname project admins are using to calculate TMRCAs. Such characteristics include the relevance of the number of testers descended from some shared SNP characterising the their Most Recent Common Ancestor (MRCA),[56] the frequencies and mean values of the relevant SNP counts, identifying which (if any) Probability Distribution Function (PDF) gives a "good fit" with these counts, and deriving associated Confidence Intervals (CIs).

This Appendix addresses findings from 17 samples of SNP counts, each with >15 testers sharing descent from a MRCA SNP, which are thought to constitute the first such survey of this subject.[57]

It is necessary to start by recognising that counts of SNPs of men sharing patrilineal descent from a MRCA SNP are but a sample, typically growing over time as more descendants test, of a larger, otherwise untested, population of a particular surname branch or haplogroup. As such the sample is subject to the mathematics of sample probabilities. It is also necessary to recognise that few genealogists are comfortable with the jargon, theoretical aspects and practical complexities of this discipline, even if many still expect that TMRCAs can somehow contribute to their own studies.

In this context it is convenient to first address the relevance of the number of samples I have surveyed and of sample size, i.e. the number of testers within each sample. Common sense tells us that where the number of samples or of testers is small then the statistical uncertainties will be dominant, and conversely that the larger the number of samples or testers available the more reliable our findings can be. In the context of YDNA SNPs, the number of samples in the public domain is small because the SNP counts of testers who share some MRCA SNP requires awareness of what FTDNA term as Private Variants, and this information is only readily available to the individual tester and his project administrator.[58] This is one reason why this subject has remained unexplored hitherto, and why in order to get most of the samples in this survey I had to resort to "citizen science" in my appeal of 8 September 2021 to the Facebook Group "Project Administrators Only". The resulting survey of 17 samples, though small, is illuminating.

Similarly the number of testers within any given sample is often small because Direct-To-Customer (DTC) Next Generation Sequencing (NGS) testing, such as FTDNA's BigY tests, is still relatively new and expensive, and few such testers can yet be grouped as sharing a MRCA SNP within the surname era.[59] Nevertheless determining the impact of the sample size, i.e. the number of testers, on the Confidence Intervals we need for TMRCA calculations is relatively straightforward, and can be illustrated by the following adaptation of Student's "t" factors:[60]

No. of testers (N)		2	3	4	5	6	7	8	9	10	12	15	20	30	50	100	∞
1 sided 84%	2-sided 68.3%	1.85	1.32	1.2	1.14	1.11	1.09	1.08	1.07	1.06	1.05	1.04	1.03	1.02	1.01	1.01	1.00
1 sided 98%	2-sided 95.4%	13.8	4.50	3.29	2.86	2.64	2.51	2.42	2.36	2.31	2.25	2.19	2.14	2.08	2.05	2.02	2.00

This table shows how the contribution of the CIs of the mean SNP count to TMRCA calculations can be modified by a simple factor to take account of the number of testers.

---

[56] This may be a single SNP, or the youngest SNP within a "block" of SNPs even if the sequence of SNPs within the block, albeit the identity of this youngest SNP is not yet known.

[57] Readers should be aware that I did not devise this survey to imitate, verify or criticise the TMRCA model developed in McDonald's 2021 paper, and I conducted the survey before I had appreciated the inherent biases in data of this type due to features such as the asymmetry of haplotrees, population size etc.

[58] The data in Alex Williamson's excellent https://www.ytree.net does not readily show the type of test (e.g. BigY500 vs. BigY700) for each tester sharing some MRCA, or the number of callable Private SNPs, or, more critically, do his haplotree branches all terminate at the same SNP level. Nor is the even more comprehensive YDNA Warehouse data available in a suitable format, though this may change with its pending move to a new platform.

[59] This Appendix is only addressing SNP counts of DTC NGS tests, and not of ancient DNA samples, or the application of SNP counts to ethnicity or haplogroup studies.

[60] Conventional "t" tables are entered with the number of "degrees of freedom", which equal N-1 where N is the sample size; hence here the number of testers must be 2 or more.

In theory these "t factors" should only be applied to Normal and Log-Normal PDFs and not to asymmetric PDFs such as Poisson, but even if the precise quantifications are not applicable to such PDFs, the underlying principle clearly applies,[61] particularly in the context of 68% CIs.[62] These considerations also explain why, in practice, as the number of testers in a sample slowly grows, the relevance of sample size decreases, and we can focus on other, more fundamental, parameters.

Before analysing the implications of these 17 samples of SNP counts it is first appropriate to develop some hypotheses. After sample size the next most significant parameter is apparently the mean of the SNP counts in each sample. If the SNP count is truly random and its mean is small, statistical theory suggests we should expect an asymmetric PDF with its mean skewed to the left such as the Poisson PDF, and if the mean count is larger we should expect Normal (aka Gaussian) PDF. Hopefully the empirical data will show us what constitutes "small" and "larger". And as the number of testers within each sample increases so the associated bar chart should become "smoother"/less "spiky", i.e. the discrete SNP counts should tend towards a smoother continuous curve, and if not then this could imply that two or more random variables which are not independent of one another are influencing the SNP counts. In such circumstances neither Poisson nor Normal PDFs will be applicable, since both cater only for a single variable, and instead some other PDF might be applicable such as the Log-Normal PDF. This PDF results from the multiplicative product of two or more independent random variables.[63] Such variables might include the size of the population, the size of the family, the age of the fathers, and/or the varying coverage of individual tests – features alluded to in McDonald's paper.[64] The Gamma PDF might also be relevant.

This hypothetical expectation of the dominant PDF of SNP count samples may be summarised thus:

Mean SNP count	expected bar chart	< 15 testers	> 15 testers	
		spiky	spiky	smooth
low	skewed to left	modified Poisson	-	Poisson PDF
high	symmetrical	Normal PDF + "t"	-	Normal PDF
any	bi-modal/spiky	Log-Normal PDF		-

We can now turn to the empirical data I have collected in the 17 samples of BigY SNP counts of testers who share a MRCA SNP. My objectives in collecting and analysing this data were:

1. To create a bar chart of the SNP count frequencies and calculate the mean, variance and standard deviation (SD) for each sample.
2. To chi-square test each sample against the Normal, Log_{10}-Normal, Poisson (based on both Mean and Variance, as in practice these input assumptions give different results) and Gamma PDFs, and thus identify the "best fit" PDF for each sample.
3. To determine the 95.4% and 68.3% CIs relevant to these mean SNP counts using three methods:
   (a) CIs derived from the relevant "best fit" PDF. For Normal and Log-Normal PDFs these are the mean +/- SD (for 68.3% CIs) or mean +/- 2*SD (for 95.4% CIs); for the Poisson PDFs I used asymmetric CIs from www.statology.org/poisson-confidence-interval-calculator.
   (b) CIs derived direct from sample data, by manual interpolation of the actual cumulative frequency probabilities of the two tails of each sample.
   (c) CIs derived from the simple, empirically derived formulae developed in section 3.4 above that are readily usable with small samples, and by project admins who are averse to mathematics.

[61] This is even more relevant when seeking the CIs of the product of two PDFs, one or both of which may be asymmetric.

[62] Alternatively a Poisson PDF can be made symmetric by use of a Log-Normal PDF.

[63] Definition from https://en.wikipedia.org/wiki/Log-normal_distribution.

[64] See McDonald 2021, 2-4, 11-12, 23. Other variables might be mutation rates changing over time or the "Founder effect", or biases in the sample, such as testers being dominantly from USA, and/or having a higher-than-average disposable income, or having been selected because they were close relatives (as opposed to being random self-selected testers). I accept the argument that the relatively large sizes of 18th and 19th century American families may have led to larger SNP counts, but I see no reason why such biases should affect the type of dominant PDF.

Methodology. For each of the 17 samples[65] I have compiled 1 or 2 sheets containing

(1) Sample data (SNP counts & frequencies), & calculations of mean, variance, Standard Deviation;
(2) Log_{10}-Normal frequencies, with associated mean, variance, Standard Deviation (SD);
(3) CIs interpolated from sample cumulative frequencies;
(4) Bar chart;
(5) Chi-squared tests against Normal, Log_{10}-Normal, Poisson and Gamma PDFs (on Sheet 2 for the larger samples) showing the results of these chi-squared test results in two forms: (i) the Excel CHISQ.TEST function probability, and (ii) the $\sum(\text{actual-predicted})^2/(\text{predicted})$ ratio;
(6) Summary

I then prepared a Summary (see Appendix B below) of my analyses of these 17 samples, listed in order of the increasing number of testers. The key to columns in this Appendix is thus:

A,B,C: Details of the sample;
D: My source (mostly e-mail references);
E: The number of testers in the sample;
F: top: The mean of the sample; 2nd line: Log_{10}-Normal mean;
G: top: Variance; bottom: "smoothness" of the bar chart (subjective);
H: top: SD; 2nd line: Log_{10}-Normal SD;
I: mid: α (for Gamma PDFs); bottom: modified Student t factor (for simplified CI estimate);
J: mid: β (for Gamma PDFs); bottom: square root of mean (for simplified CI estimate);
K: 1st five lines: the 5 PDF types against which I have tested each sample, and derived CIs;
Last two lines: 2 additional methods I have used to derive CIs;
L: Excel CHISQ.TEST results copied from individual data sheets (0 = poor fit, 1 = good fit);
M: sums of $(\text{actual-predicted})^2/(\text{predicted})$ similarly copied (low=good fit, high=poor fit);
For L & M **bold font** indicates the PDF that gives the best fit to the sample data;
N: 95.4% CIs calculated from PDF types and from cumulative frequencies;
O: 68.3% CIs calculated from PDF types, from cumulative frequencies, and from simplified estimate;
For N & O shaded background indicates use of PDF best fits inappropriate as they are < [0.80];
bold font indicates cumulative frequency CIs which are wider than the PDF best fit;
italic font indicates the simplified estimate CIs which are narrower than the relevant PDF best fit or the cumulative frequency CIs;
P: cameo of bar chart copied from the individual data sheet.

Problems. In this analysis of the 17 samples I had to address a number of problems:

- The processes of data acquisition and analysis were tedious and error-prone, even when using Excel functions, because of:
  - possible inconsistent determination of callable, phylogenetically-significant SNPs;[66]
  - random errors in manually identifying or counting of these SNPs;
  - possible systemic errors in my analyses (especially in log-normal PDFs & chi-squared tests);
  - random errors in developing Excel data sheets (volume of data makes processing error-prone);
  - misinterpretation of the calculations.
- Neither the Excel CHISQ.TEST critical values nor the $\sum(\text{actual-predicted})^2/\text{predicted}$ parameter seem adequate alone to identify the "best fit", and so I took account of both indices.
- While for a "pure" Poisson PDF, the mean is equal to the variance, in practice either can be used to characterise the Poisson PDF against which the sample fit is compared.
- Where the variance-based Poisson PDFs were better (i.e. the R1a and Strother samples) I also tried averaging the mean and variance, but these attempts did not yield better fits.[67]

---

[65] All the 17 samples are counts of SNPs, private and intermediate, of testers sharing some MRCA SNP. Data sources are cited on the individual data sheets. Note that I normalised the "Royal Stewart" data to a common 10M bps length, and the O'Brian R-DC782 data counts the SNPs back to R-L226, not to the DC782 MRCA.
The data for the Border Irwin R-FGC13746 sample is included as an example in Appendix C. The data sheets for the other 16 samples are available on request.

[66] Note it is the consistency of the SNP counts within each sample that matters, not the consistency between the samples.

- For many samples I was unable to calculate Gamma PDF data, though this doesn't appear critical.
- The optimum method of calculating CIs of Poisson PDFs is contentious: at least 20 different methods can be considered.[68] I have used the asymmetric upper and lower CIs derived from the convenient www.statology.org/poisson-confidence-interval-calculator, which according to www.statology.org/poisson-confidence-interval are $0.5*X^2_{2N,\ \alpha/2}$ and $0.5*X^2_{2(N+1),\ 1-\alpha/2}$ respectively, where X^2 is the chi-squared critical value, N is the number of "Observed Events" (in this application the mean SNP count), and α is the significance level (in this application either 95.4% or 68.3%).
- The CIs derived direct from the sample cumulative frequencies of SNP counts (as opposed to those derived from theoretical PDF models) are sensitive to outliers, have to be interpolated manually, and hence lack reliability; they are inappropriate for 95% CIs and for small samples.

Findings. The following findings emerging from this exercise seem appropriate:[69]

1. Samples of less than about 15 testers sharing some MRCA SNP are of little use in establishing their statistical characteristics.[70] There is thus some merit in developing a simple practical tool, based on analyses of larger samples, for guiding project admins of surname projects with small samples on developing indicative Confidence Intervals for TMRCA estimates based on the convenient mean SNP counts; such a tool may be of some use to admins averse to mathematics.
2. All of 17 samples with more than 15 testers are samples of the larger, real-world populations of the relevant surname/haplotree branch. Each sample is growing over time as more members take NGS tests.
3. Contrary to my expectations, in practice none of the samples show a close fit to their associated Normal, Log_{10}-Normal, Poisson or Gamma PDFs.[71]
4. The best fit I was able to find was with the Border Irwin R-FGC13746 Y700 sample. Ironically here there was little to choose between the closeness for the fits with its associated Normal, Log_{10}-Normal and Poisson (mean) PDFs.
5. Of the 17 samples:
   - 9 have a "best fit" with their Poisson (mean) PDFs.
   - 2 have a "best fit" with their Poisson (variance) PDFs: R1a (poor data, poor fit) and Strother.
   - 4 have a "best fit" with their Normal PDFs: Akins AF06 (data is skewed to the right), O'Brian (ancestral SNP L226 lived over 1,500 years ago), R-L226 Y700 (likewise) and Doherty Y500; in all four their Poisson (mean) PDFs were a "second best fit".
   - 2 have a "best fit" with their Log_{10}-Normal PDFs: R-L513 Y500 and R-L513 Y700; however neither are good fits and both are only marginally better fits than with their Normal PDFs.
6. These findings, albeit inconclusive, can be interpreted several ways. It seems apparent that:
   - for chi-squared and CI calculations based on Poisson PDFs it is preferable to use the mean rather than the variance;[72]

---

[67] Ralph Taylor found that for the Border Irwin R-FGC13746 data using x=n-2 gave marginally better fits, but my applying similar methods with other samples failed to reproduce any improvements and I have not pursued this sophistication.

[68] See www.ine.pt/revstat/pdf/rs120203.pdf. The Statology algorithm is apparently adapted from the Garwood method.

[69] It is curious that where the mean SNP counts of Y500 and Y700 data can be compared in the samples addressed in Appendix B, the Y700/Y500 ratios are 1.19 (Doherty), 1.15 (Border Irwin), 1.10 (R-L513) and 1.20 (R-L226), whereas the SNP mutation rates of derived from the YDNA Warehouse database have a Y700/Y500 ratio of 1.57.

[70] An extreme example was offered to me by Mary Wiley for eight I-Y20863 BigY500 testers whose SNP counts were: 2: 4; 3: 1; 4: 1: 5:0; 6:2: a "U" shaped distribution rather than the typical "Π" shaped distribution!

[71] I have not attempted fits of Log_e-Normal, Exponential or Binomial PDFs, but the "spikiness" of the samples suggests that close fits with any common PDF are unlikely.

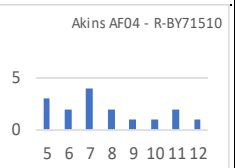
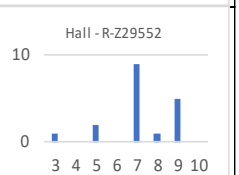
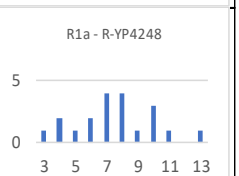
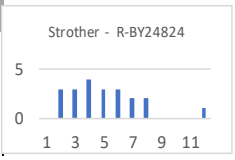
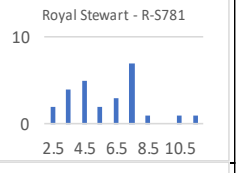
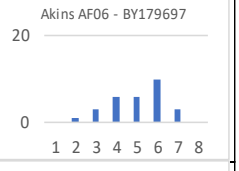
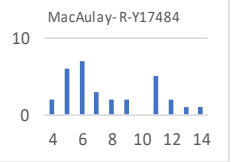
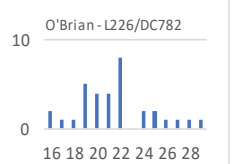
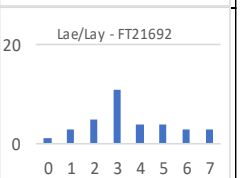
[72] That said, the Poisson PDF is not a "good fit" to every such SNP count. A measure of the relevance of Poisson PDFs to the sample data is the ratio of the variance to the mean, a ratio of 1.00 indicating the sample fits a Poisson PDF exactly:
Variance/mean = 0.3-0.6: Akins (x2), Hall, O'Brian;
Variance/mean = 0.7-0.9: R1a, Royal Stewart, Lae/Lay, Little, Doherty (x2), Border Irwin, R-L513 (x2), R-L226 (x2);
Variance/mean = 1.1-1-2: Strother, MacAulay.

- for samples with mean SNP counts of < c.15, i.e. those for TMRCA calculations within the surname era, the Poisson PDF is the dominant "best fit" PDF;
- for samples with mean SNP counts > c.15, i.e. those used for TMRCA calculations for ethnicity and haplogroup the dominant PDF is less clear.

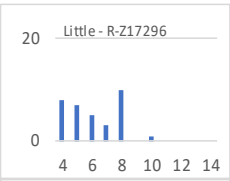
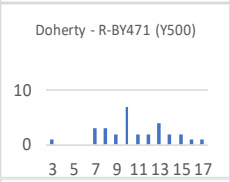
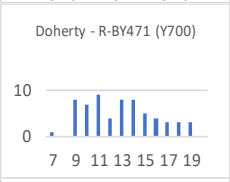
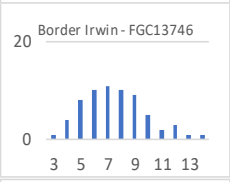
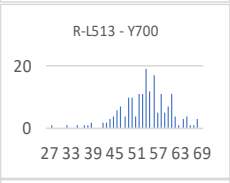
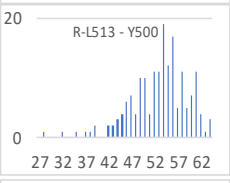
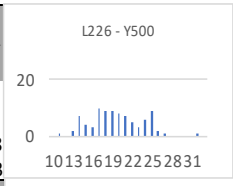
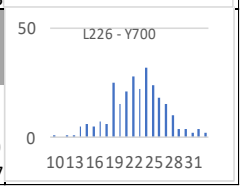
7. The failure of this survey to identify a robust a single "best fit" PDF for all the 17 samples does not invalidate my quest for some indicative CIs relevant to TMRCAs estimated from the convenient mean of SNP counts, even though these TMRCA estimates have been shown to be less accurate than when estimated using the unbiased node-by-node method. As explained above CIs can also be derived be three means: from "best fit" PDFs, from the cumulative frequencies of the sample SNP counts, and by the empirically derived formulae or table developed in section 3.4 above. These CIs are shown in columns N and O of Appendix B, with **bold font** indicating cumulative frequency CIs which are wider than the best fit PDF CIs and *italic font* indicating the empirically derived CIs which are narrower than the relevant best fit PDF CIs or of the cumulative frequency CIs. It will be seen that for curious genealogists, for those averse to lengthy calculations, and for samples of less that c.15 testers, the formulae developed in section 3.4 above offer indicative CIs that are "safe" but conservative.

8. However there is a more fundamental lesson arising from this survey. With the possible exception of the Border Irwin sample, there is no evidence to support my expectations that larger samples (e.g. the R-L513 and R-L226 Y500 samples and the Doherty, R-L513 and R-L226 Y700 samples) are tending towards either a smoother bar chart or a close fit with one of the common PDFs. This conclusion is consistent my belated recognition that while SNP mutations are random, the downstream counts of the SNPs of individuals sharing a MCRA are not random, but are instead biased because of the asymmetric shape of most haplotrees, and possibly other factors such as the impact of changing population size, varying family size, age of the father, Founder effect etc. At present so little is understood about such variables that we cannot even speculate on how they affect SNP counts. These considerations also suggest a Bayesian approach such as that advocated by McDonald, but this is beyond the scope of this paper.

Appendix B - Summary of analyses of 17 samples of SNP counts - 1

A B C D	E	F	G	H	I	J	K	L	M	N	O	P
Data set	No.	Mean	Var-	SD	α	β	PDF type	CHISQ.	Σ	Confidence intervals		Bar chart
Ancestral test	of		iance		(Gamma PDFs)		(for chi.sq test & CIs)	TEST	diff.^{2}/f	95.4%	68.3%	
SNP type source	testers		bar chart		TF_L	sq.rt. mean		(Excel)				
Akins AF04	16	7.75	4.81	2.19			Normal	0.15	18	3.0 - 12.3	5.3 - 10.0	
R-BY1510		7.45		1.02			Log_{10}-Normal	0	3637	5.2 - 9.3	6.2 - 8.3	
Y700							**Poisson on mean**	**0.91**	**7**	3.2 - 15.5	5.0 - 11.6	
							Poisson on variance	0	72			
					12.5	1.61	Gamma					
			fair				interpolated direct			4.1 - 12.7	**4.8** - 11.2	
MC					1.04	2.78	Simplifed estimate				*4.9* - 12.0	
Hall/Smith	18	7.17	2.47	1.57			Normal	0.03	**21**	4.0 - 10.3	5.6 - 8.7	
R-Z29552		6.24		1.06			Log_{10}-Normal	0	149	4.1 - 8.4	5.2 - 7.3	
Y700							**Poisson on mean**	**0.08**	29	2.9 - 14.3	4.5 - 11.1	
							Poisson on variance	0	2224			
					20.8	2.90	Gamma	0	-			
			poor				interpolated direct			**2.5** - 9.9	4.9 - 9.5	
DH					1.03	2.75	Simplifed estimate				4.3 - 12.4	
R1a	20	7.55	6.05	2.46			Normal	0	311	2.4 - 12.3	5.1 - 10.0	
R-YP4248		7.08		1.03			Log_{10}-Normal	0	huge	5.0 - 10.1	6.0 - 8.1	
Y700							Poisson on mean	0	68			
							Poisson on variance	**0.01**	**27**	3.1 - 15.5	4.8 - 11.4	
					9.43	2.46	Gamma					
			poor				interpolated direct			**2.5** - 13.5	**4.1** - 10.7	
MC					1.03	2.25	Simplifed estimate				*5.2 - 11.0*	
Strother	21	5.05	5.76	2.40			Normal	0.04	22	0.2 - 9.9	2.6 - 7.5	
R-BY24824		4.7		1.02			Log_{10}-Normal	0	125	2.7 - 9.9	2.6 - 7.5	
Y700							Poisson on mean	0.33	15			
							Poisson on variance	**0.62**	**11**	1.6 - 11.8	2.9 - 8.5	
							Poisson on mean/var	0.59	11			
					4.42	0.88	Gamma	0	66			
			good				interpolated direct			**1.1 - 17.5**	3.2 - 7.9	
MC					1.03	2.25	Simplifed estimate				2.7 - 9.5	
Royal Stewart	26	5.96	5.17	2.27			Normal	0	775	1.4 - 10.5	3.7 - 8.2	
R-S781		5.52		1.03			Log_{10}-Normal	0	61	3.5 - 7.6	4.5 - 6.5	
mixed							**Poisson on mean**	**0.80**	**11**	2.1 - 13.2	3.6 - 9.6	
(normalised							Poisson on variance	0.63	13			
to 10M bps)					6.87	1.15	Gamma	0.15	25			
			poor				interpolated direct			**1.9** - 11.9	**3.0** - 8.3	
IM					1.02	2.44	Simplifed estimate				3.5 - 9.7	
Akins AF06	29	5.03	1.69	1.3			**Normal**	**0.93**	**4**	2.3 - 7.6	3.6 - 5.2	
R-BY179697		4.84		1.02			Log_{10}-Normal	0.13	11	2.7 - 6.7	3.7 - 5.7	
Y700							Poisson on mean	0.36	14	1.6 - 11.8	2.9 - 8.5	
							Poisson on variance	0	844			
					15.01	2.98	Gamma		?			
			good				interpolated direct			**1.7 - 7.7**	**3.1 - 6.8**	
MC					1.02	2.24	Simplifed estimate				2.7 - 8.5	
MacAuley	32	*7.59	8.99	3.00			Normal	0.40	19	1.6 - 13.6	4.6 - 10.6	
R-Y17484		7.00		1.04			Log_{10}-Normal	-	huge	4.8 - 9.2	5.9 - 8.1	
Y700							**Poisson on mean**	**0.56**	**18**	3.1 - 15.3	4.8 - 11.4	
		*: mean of					Poisson on variance	0.16	29			
		Y29170			6.41	0.85	Gamma	?	114			
		= 11.7	fair				interpolated direct			**1.7** - 14.5	**4.3 - 11.7**	
KM					1.02	2.75	Simplifed estimate				*4.8* - 11.8	
O'Brian DC782	33	21.51	9.76	3.12			**Normal**	**0.88**	**15**	15.2 - 27.8	18.4 - 26.4	
R-L226		21.28		1.00			Log_{10}-Normal	0	huge	19.3 - 23.3	20.3 - 22.3	
Y700							Poisson on mean	0.63	22	13.3 - 32.9	16.9 - 27.2	
(filtered							Poisson on variance	0	huge			
by DW)					47.4	2.20	Gamma		?			
			fair				interpolated direct			16.4 - **29.2**	**18.2** - 26.4	
DO/RC					1.02	4.64	Simplifed estimate				16.8 - 28.6	
Lae/Lay	34	3.56	3.25	1.80			Normal	0.785	7	0 - 7.2	1.8 - 5.4	
R-FT21692		2.92		1.16			Log_{10}-Normal	0.916	9	0.5 - 5.7	1.8 - 4.1	
Y700							**Poisson on mean**	**0.999**	**6**	0.9 - 9.7	1.8 - 6.6	
							Poisson on variance	0.997	8			
					3.9	1.06	Gamma	?	6			
			good				interpolated direct			**0** - 7.7	**1.3** - 6.1	
MG					1.02	1.89	Simplifed estimate				*1.6 - 6.5*	

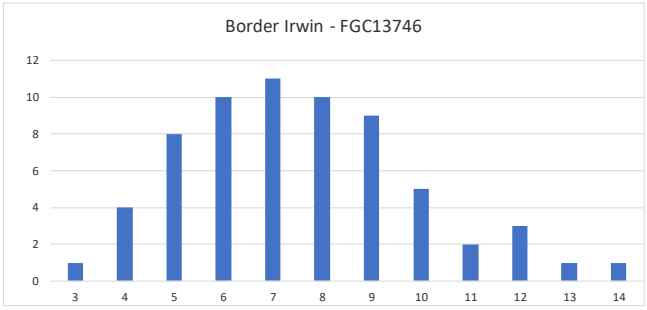
Appendix B - Summary of analyses of 17 samples of SNP counts - 2

A B C D	E	F	G	H	I	J	K	L	M	N	O	P
Data set Ancestral SNP / test type / source	No. of testers	Mean	Var-iance / bar chart	SD	α (Gamma PDFs) TF_L	β (Gamma PDFs) sq.rt. mean	PDF type (for chi.sq test & CIs)	CHISQ. TEST (Excel)	∑ diff.²/f	Confidence intervals 95.4%	Confidence intervals 68.3%	Bar chart
Little	37	5.81	3.78	1.94			Normal	0.15	18	1.9 - 9.7	3.9 - 7.8	
R-Z17296		5.51		1.04			Log_{10}-Normal	0	1904	3.4 - 7.6	4.5 - 6.5	
Y700							**Poisson on mean**	**0.2**	20	2.0 - 12.9	3.5 - 9.4	
							Poisson on variance	0	115			
					8.94	1.54	Gamma	0	89k			
			poor				interpolated direct			**0.9** - 10.1	**3.3** - 8.4	
JL/TL					1.02	2.41	Simplifed estimate				3.3 - 9.5	
Doherty	30	10.83	9.41	3.07			**Normal**	**0.90**	**14**	4.7 - 17.0	7.8 - 13.9	
R-BY471		10.30		1.02			Log_{10}-Normal	0	huge	8.3 - 12.3	9.3 - 11.3	
Y500							Poisson on mean	0.83	16	5.2 - 19.5	7.2 - 15.2	
							Poisson on variance	0.51	19			
					12.5	1.15	Gamma	?	1100			
			fair				interpolated direct			**2.7** - 17.3	7.2 - 14.7	
ZD					1.02	3.29	Simplifed estimate				*7.5* - 15.9	
Doherty	63	12.90	8.91	2.99			Normal	0.73	18	6.9 - 18.9	9.9 - 15.9	
R-BY471		12.59		1.01			Log_{10}-Normal	0	6651	4.5 - 20.7	8.6 - 16.6	
Y700							**Poisson on mean**	**0.84**	15	6.7 - 22.3	9.3 - 17.6	
							Poisson on variance	0	265			
					10.7	1.45	Gamma	?0	**9**			
			fair				interpolated direct			8.1 - 19.5	**9.1** - 16.8	
ZD					1.02	3.59	Simplifed estimate				*9.3* - 18.3	
Border Irwin	65	7.57	5.48	2.34			Normal	0.994	6	2.9 - 12.2	5.2 - 10.9	
R-FGC13746		7.21		2.39			Log_{10}-Normal	0.9994	8	5.3 - 9.0	6.3 - 8.1	
Y700							**Poisson on mean**	**0.9999**	**5**	3.1 - 15.3	4.9 - 11.5	
							Poisson on variance	0	70			
					10.44	1.38	Gamma	0	907			
			good				interpolated direct			3.1 - 13.5	**4.6** - 10.4	
JI					1.01	2.75	Simplifed estimate				*4.8* - 11.7	
R-L513	185	53.64	44.63	6.68			Normal	0	199	40.3 - 67.0	47.0 - 60.3	
Y700		53.21		6.70			**Log_{10}-Normal**	0	**163**	41.8 - 68.6	48.5 - 61.9	
							Poisson on mean	0	178	40.0 - 70.3	46.4 - 62.0	
							Poisson on variance	0	52k			
					64.5	1.2	Gamma	-	huge			
			v.spiky				interpolated direct			**37.3** - 66.7	46.7 - 60.6	
DV					1.00	7.34	Simplifed estimate				46.3 - 64.6	
R-L513	237	48.78	43.16	6.57			Normal	0	90	35.6 - 61.9	42.2 - 55.3	
Y500		48.31		6.59			**Log_{10}-Normal**	0	**83**	35.1 - 61.5	41.7 - 54.9	
							Poisson on mean	0	93	35.8 - 65.6	41.8 - 55.7	
							Poisson on variance	0	330			
					55.2	1.13	Gamma	0	50k			
			v.spiky				interpolated direct			**32.5** - 61.3	42.0 - 55.7	
DV					1.00	6.98	Simplifed estimate				41.8 - 59.2	
R-L226	87	19.68	15.46	3.94			Normal	0.06	37	11.8 - 27.5	14.8 - 23.6	
Y500 (filtered by DW)		19.47		3.94			Log_{10}-Normal	0.05	40	11.6 - 27.3	15.5 - 23.4	
							Poisson on mean	**0.66**	**26**	11.8 - 30.7	15.3 - 25.2	
							Poisson on variance	0	25k			
					25.1	1.27	Gamma	0	6575			
			fair				interpolated direct			12.5 - 27.0	14.9 - **25.8**	
DO/RC					1.01	4.35	Simplifed estimate				*15.3* - 26.3	
R-L226	262	23.55	20.85	4.57			**Normal**	**0.38**	**38**	19.0 - 28.1	14.4 - 32.7	
Y700 (filtered by DW)		25.47		5.37			Log_{10}-Normal	0	108	14.7 - 36.2	20.1 - 30.8	
			28				Poisson on mean	0.30	39	14.9 - 35.3	18.7 - 29.5	
							Poisson on variance	0	231			
					26.6	1.13	Gamma	0	3070			
			good				interpolated direct			**13.6** - **32.8**	17.5 - 27.0	
DO/RC					1.00	4.85	Simplifed estimate				*18.7* - 30.7	

Appendix C - Border Irwin FGC13746 Y700 SNP counts

(1) Sample data						(2) Log_{10}-Normal calculation					(3) CIs direct from sample data		
c	f	fc	c-mean	$(c\text{-mean})^2$	$f(c\text{-mean})^2$	$log_{10}c$	$f*log_{10}c$	c-mean	$(c\text{-mean})^2$	$f*(c\text{-mean})^2$	f% of ∑f	cum f%s	
3	1	3.00	-4.57	20.88	20.88	*0.477*	*0.477*	*0.477*	*0.228*	*0.228*	1.54%	1.54%	$LCI_{95.4\%}$ @ 2.3% = c. 3.1
4	4	16.00	-3.57	12.74	50.96	*0.602*	*2.408*	*0.602*	*0.362*	*1.450*	6.15%	7.69%	
5	8	40.00	-2.57	6.60	52.81	*0.699*	*5.592*	*0.699*	*0.489*	*3.909*	12.31%	20.00%	$LCI_{68.4\%}$ @15.6% = c. 4.6
6	10	60.00	-1.57	2.46	24.62	*0.778*	*7.780*	*0.778*	*0.605*	*6.053*	15.38%	35.38%	
7	11	77.00	-0.57	0.32	3.56	*0.845*	*9.295*	*0.845*	*0.714*	*7.854*	16.92%		
8	10	80.00	0.43	0.19	1.86	*0.903*	*9.030*	*0.903*	*0.815*	*8.154*	15.38%		
9	9	81.00	1.43	2.05	18.42	*0.954*	*8.586*	*0.954*	*0.910*	*8.191*	13.85%		
10	5	50.00	2.43	5.91	29.54	*1.000*	*5.000*	*1.000*	*1.000*	*5.000*	7.69%	18.46%	$UCI_{68.4\%}$ @15.6% = c.10.4
11	2	22.00	3.43	11.77	23.54	*1.041*	*2.082*	*1.041*	*1.084*	*2.167*	3.08%	10.77%	
12	3	36.00	4.43	19.63	58.90	*1.079*	*3.237*	*1.079*	*1.164*	*3.493*	4.62%	7.69%	
13	1	13.00	5.43	29.49	29.49	*1.114*	*1.114*	*1.114*	*1.241*	*1.241*	1.54%	3.08%	$UCI_{95.4\%}$ @ 2.3% = c.13.5
14	1	14.00	6.43	41.35	41.35	*1.146*	*1.146*	*1.146*	*1.313*	*1.313*	1.54%	1.54%	
Sum	65	492.00			355.94		*55.747*			*49.053*			
Mean = ∑fc/∑f =					7.57					*0.858*			
										7.21			
Variance = $∑f(c\text{-mean})^2/∑f$ =					5.48					*0.755*			
										5.69			
Normal PDFs:			SD = sqrt.variance =		2.34					2.39			
Gamma PDFs:			Mean = α/β		Var'n = $α/β^2$								
			α =		10.45								
			β =		1.38								

(4) Bar chart:



(5) Chi^2 tests

		Normal		**Poisson**		**Poisson**		**Log10-Normal**		**Gamma**	
		mean =	**7.57**	**mean =**	**7.57**	**variance=**	**5.48**	**mean =**	**7.21**	α =	10.44
		Std. Dev'ı	**2.34**					**Std. Dev**	**2.39**	β =	1.38
c	f, actual	f predicted	$diff^2/f$	f predicted	$diff^2/f$	f predicted	$diff^2/f$	f predicted	$diff^2/f$	f predicted	$diff^2/f$
0	0	*0.08*	0.08	0.03	0.03	0.27	0.27	*0.15*	0.15	0.00	
1	0	0.22	0.22	0.25	0.25	1.49	1.49	0.37	0.37	0.00	
2	0	0.65	0.65	0.96	0.96	4.07	4.07	1.01	1.01	0.00	
3	1	1.65	0.25	2.42	0.84	7.43	5.57	2.30	0.73	0.01	118.77
4	4	3.46	0.08	4.59	0.07	10.18	3.75	4.40	0.04	0.06	255.89
5	8	6.06	0.62	6.94	0.16	11.16	0.90	7.08	0.12	0.24	249.25
6	10	8.85	0.15	8.76	0.18	10.19	0.00	9.54	0.02	0.65	133.52
7	11	10.76	0.01	9.47	0.25	7.98	1.14	10.81	0.00	1.36	68.45
8	10	10.90	0.07	8.97	0.12	5.47	3.76	10.27	0.01	2.32	25.40
9	9	9.19	0.00	7.54	0.28	3.33	9.66	8.20	0.08	3.42	9.11
10	5	6.46	0.33	5.71	0.09	1.82	5.53	5.49	0.04	4.48	0.06
11	2	3.78	0.84	3.93	0.95	0.91	1.31	3.09	0.38	5.34	2.08
12	3	1.85	0.72	2.48	0.11	0.41	16.10	1.46	1.64	5.88	1.41
13	1	0.75	0.08	1.44	0.14	0.17	3.89	0.58	0.31	6.06	4.23
14	1	0.25	2.19	0.78	0.06	0.07	12.67	0.19	3.41	5.91	4.08
15	0	0.07	0.07	0.39	0.39	0.03	0.03	0.05	0.05	5.49	5.49
16	0	0.02	0.02	0.19	0.19	0.01	0.01	0.01	0.01	4.90	4.90
17	0	0.00	0.00	0.08	0.08			0.00	0.00	4.20	4.20
18	0	0.00	0.00	0.03	0.03			0.00	0.00	3.49	3.49
19	0	0.00	0.00	0.01	0.01			0.00		2.82	2.82
20	0	0.00	0.00	0.01	0.01			0.00		*8.37*	*8.37*
∑	65	65.00	6.40	65.00	5.20	65.00	70.16	65.00	8.38	65.00	901.54
Chi^2 value		0.9943		0.99991		0.000		0.9994		0.0675	

(6) Summary

Sample measures:

Mode =	7
Median = (14-3)/2 =	8.5
Mean =	7.6
Variance =	5.5
SD =	2.3

Comparisons of data with other PDFs:

Best fit: Poisson (mean) (variance/mean ratio = 0.7)

Good fits: Log_{10}-Normal Normal

Poor fits: Poisson (variance) Gamma

CIs:	95.4%	68.4%
Normal:	2.9-12.2	5.2-10.9
Log_{10}-Normal:	5.3- 9.0	6.3- 8.1
Poisson (mean):	3.1-15.3	4.9-11.5
ex cumulative frequencies:	3.1-13.5	4.6-10.4
ex equations (10) & (11):	-	4.8-11.7