Supplementary Paper No.3

IDENTIFYING SUB-GROUPS WITHIN THE BORDER IRWINS GENETIC FAMILY: TiPS, SNPS, AND CLADOGRAMS

James M Irvine, Study Administrator, April 2011, revised 9 June 2011

One of the original goals of this Study is "to identify DNA profiles of the Bonshaw and Drum branches and sub-branches thereof". Time has moved on, and this might now read better as

> *to identify the DNA signatures of the genetic families bearing [our] surnames, both now and formerly, and sub-branches thereof*.[1]

We have already identified modal DNA signatures for genetic families that include Bonshaw, Drum and a surprising number of other families that use our surname. And within our Borders genetic family, which includes two thirds of our participants, we now know the DNA signatures of representatives of the genealogical lines of Bonshaw, Castle Irvine, Dumfries and Eskdale. But hitherto we have made little progress in identifying sub-groups within this large genetic family. Ideally a combination of genetic, genealogical, demographic and historical knowledge will enable the identification of genetic sub-groups of participants who are all descended from a common ancestor who probably lived in Dumfriesshire in the 13th or 14th century. This grouping in turn should help us to identify the likely origins, both when and whence, of these various sub-groups. This paper represents a significant step in this direction. One benefit arising from this development is the increased likelihood of existing participants within our Borders family being able to identify other participants in their sub-group to whom they may be genealogically related.

Various tools are available to help identify sub-groups, but genetic genealogy is an emerging art whose rules and procedures have to be developed and adapted as we go along. Techniques presently available include:[2]

1. Visual identification of groups of participants sharing common genetic features.
2. FTDNA's TiP tool.
3. Haplogroups and SNPs.
4. Cladograms - network diagrams.
5. Cladograms - phylogenetic trees.

All these methods have now been applied to the data of this Study. All have been found to have strengths and weaknesses, which I discuss below.

1. Visual identification.

Visual perusal of tabulated test results can identify various repetitive features which suggest sub-groups. Intuitively the identification of fast mutating markers, unusual markers,[3] and known genealogical relationships should enhance the chances of identifying sub-groups. This approach has enabled some progress, but by definition it is inherently subjective and if used alone is unlikely to lead to a robust tool that will maximise the potential of our test results.

2. FTDNA's TiP tool.

Coupled with some genealogical and historical knowledge, this has proved a robust tool for dividing our Study database into about 20 genetic families.[4] But the range of 24-generation TiP probabilities within the large Borders family is so narrow that TiPs alone cannot alone resolve the challenge of identifying sub-groups therein.

3. Haplogroups and sub-clades (SNP tests).

Some surname studies use these tests to identify the age and geographic origin of genetic families. As this is "deep ancestry" and thus outside the scope of this Study I initially dismissed them, but as geneticists assure me that these are the most reliable tool for identifying genetic families I am now including summaries of the SNP test results for those participants involved in our on-going results table. I am pleased to confirm

---

[1] The semantics are confusing to many. FTDNA use the term "DNA signature" rather than "DNA profile", and I prefer "genetic family" to "branch", "group", "clade" or "cluster", and" sub-group" to "sub-clade" or "sub-cluster".

[2] Some consider use of TMRCA tables and McGee's calculator, but for my views on these see section 9.4 of the accompanying Supplementary Paper No.1 ("Towards Improvement ....").

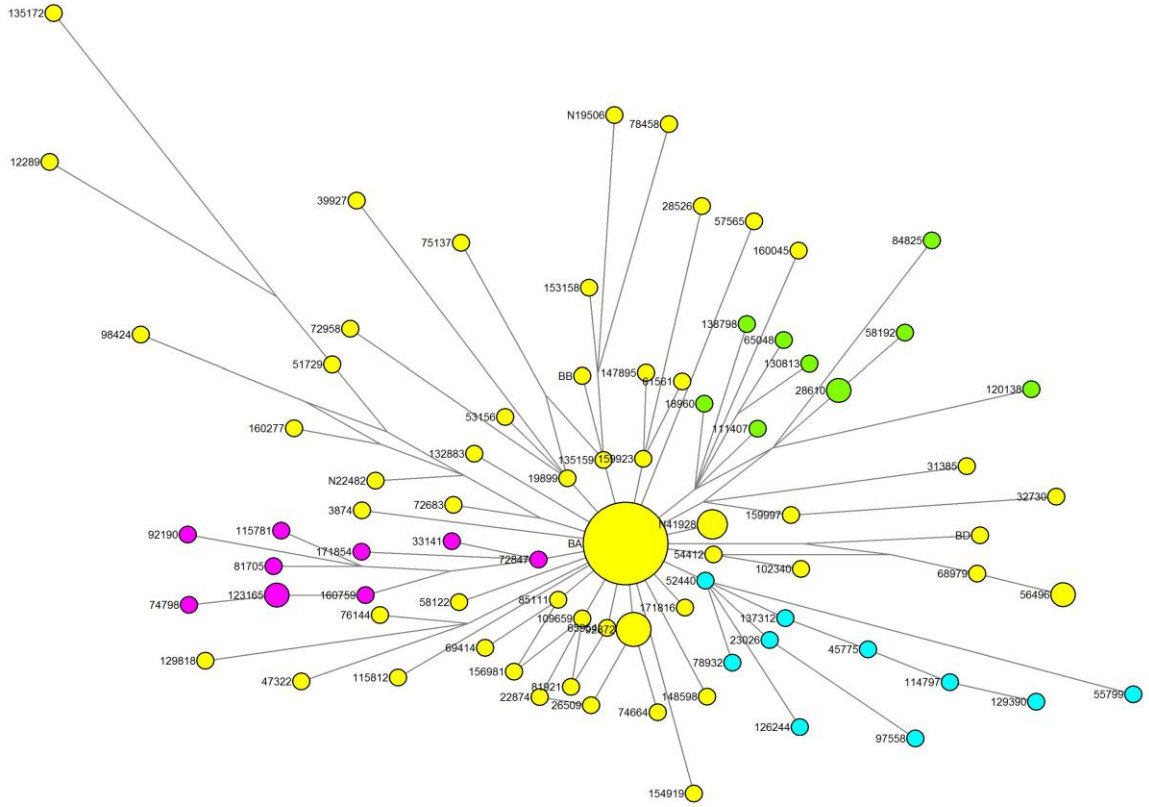
[3] See section 9.2 of the accompanying Supplementary Paper No.1 ("Towards Improvement ....").

[4] See Appendix C of the accompanying Supplementary Paper No.1 ("Towards Improvement ....")..

our existing genetic family divisions are not inconsistent therewith. For example, the M222 test has clearly distinguished our Perthshire group (PF) from other genetic families. However the interpretation of SNP data is not straightforward: FTDNA's new phylogenetic tree is not yet available on-line,[5] new SNP tests become available with increasing frequency, and it is apparent even the latest SNP tests do not yet distinguish between all genetic families, let alone sub-groups.[6] This is because at present the bifurcations identified by SNP tests occurred before the sub-groupings within our Borders family, which by definition occurred after the first adoption of our hereditary surname, which was first recorded in the 13th century or, at 30 years per generation, about 24 generations ago.

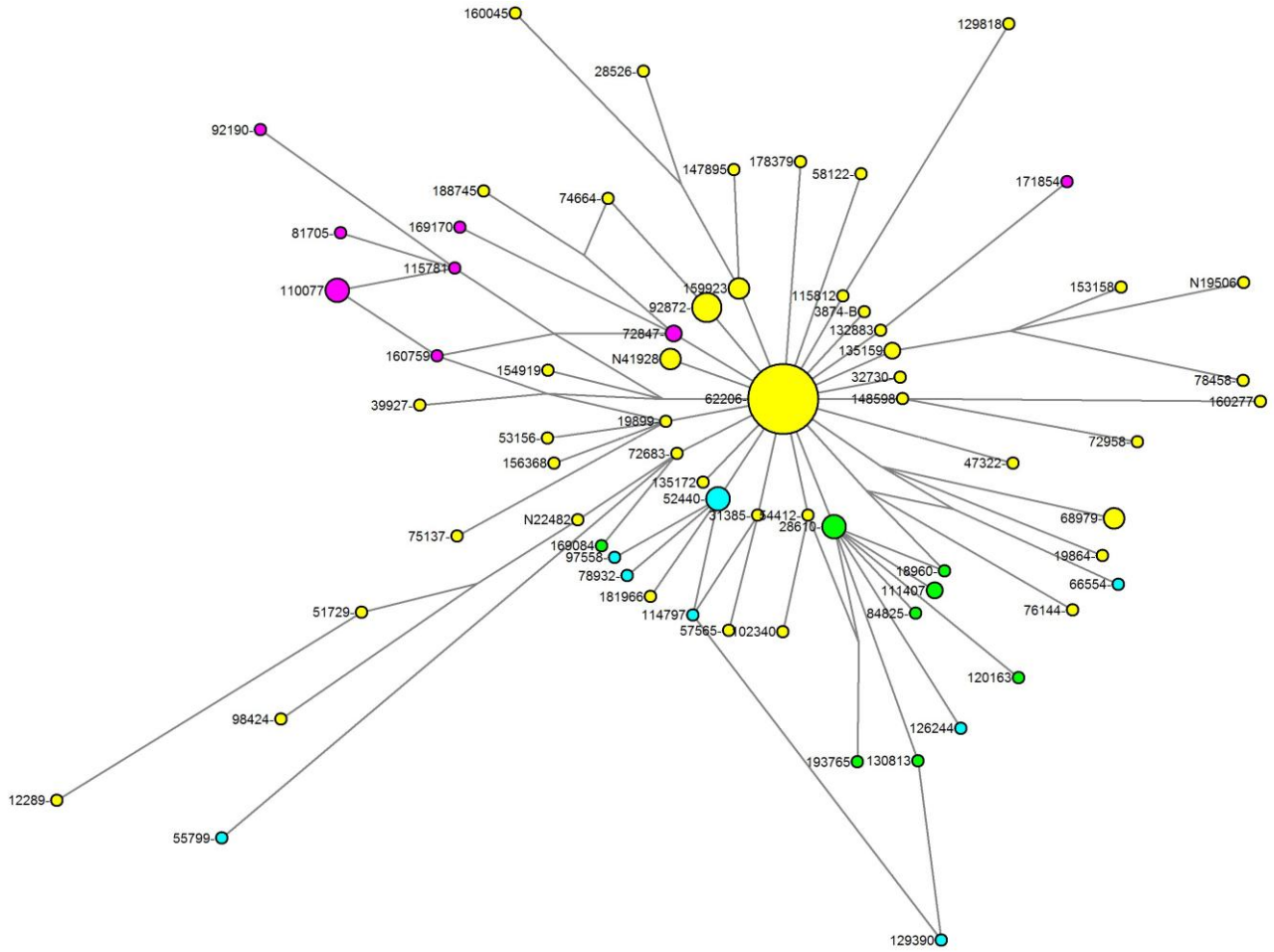
4. Cladograms - network diagrams.

A cladogram is a diagramatic representation of ancestral relationships. Network diagrams, which convey some of the ambiguity in determining these relationships, are the output of sophisticated software packagers adapted from engineering network analysis algorithms. These packages can only compare data of equal resolution, i.e. in the genetic context all data must be, for example, 37 markers or 67 markers; the packages are not easy to apply; and they require several subjective inputs such as what weightings to input for individual marker mutation rates (a contentious issue in itself). The most popular software is Fluxus, and I am grateful to one of our participants, Rick Byers, for his time and patience in applying this tool to our data. For a project of our size network diagrams are neither easy to read nor useful when applied to all our data, but when used for just our Borders genetic family and giving a weight of zero to the two fastest mutating markers CDYa & b and the four multi-copy markers DYS463a, b, c & d,[7] informative results become possible. Two outcomes illustrating the effect of different mutation rate weightings are shown below:



Network diagram of Border Irwins 37-marker data, March 2010
(yellow: B11; green: B12; blue: B2; magneta: B3)
The size of each node represents the number of participants with identical DNA signatures;
radial distances from the hub reflect genetic distance; azimuth features are arbitrary;
bifurcations without nodes imply deficiencies in the data set,[8] and convergences imply mutations that have reverted.

[5] The 2011 edition of FTDNA's 2011 Y tree may be purchased from them direct. They introduced this phylogenetic tree in March 2011, though alas their nomenclature is not consistent with the ISOGG tree. Great care is needed when referring to R1b1 nomenclature, and it is safer to refer to the relevant SNP test.
[6] For further details see the accompanying Supplementary Paper No.5 (SNP Test Results).
[7] For a similar conclusion, see www.johnbrobb.com/Content/DNA/TMRCA&GD.pdf
[8] In Appendices A and B of the accompanying Supplementary Paper No.1 ("Towards Improvement ....") show that our project has only attracted less than 0.2% of the Irwins etc. alive today.



Network diagram of Border Irwins 37-marker data, February 2011

These diagrams are suggestive of possible sub-groupings, but when used alone are not sufficient to rigorously and unambiguously define such divisions.

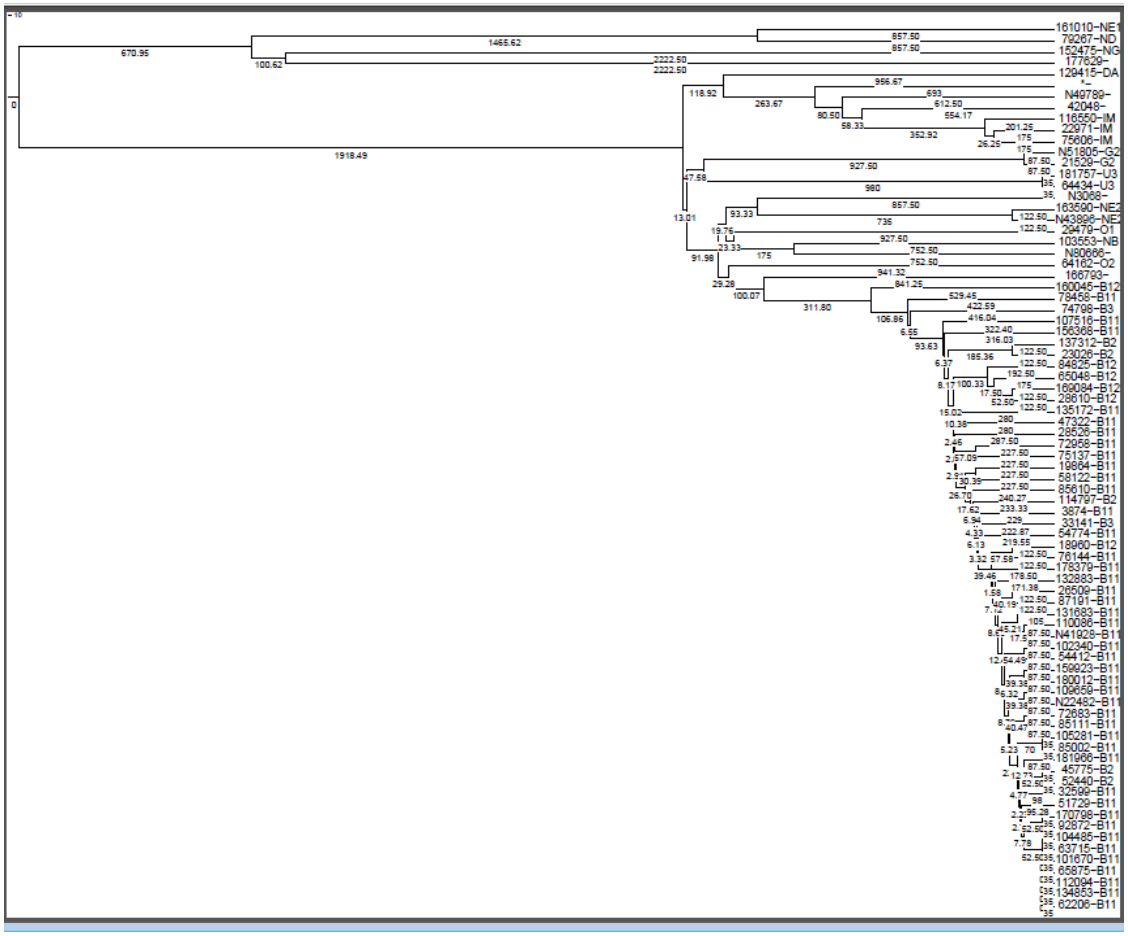
5. Cladograms - phylogenetic trees.

Like network diagrams, these are complex software packages that can only handle data of common resolution, and also need a number of subjective inputs. But instead of indicating genetic distances and convergences they introduce a time dimension for the genetic bifurcations, conventionally with their horizontal axis showing the present day on the right, receding back in time to the left. One example of a phylogenetic tree, with the timescale adjusted for optimal visual clarity, is FTDNA's depiction of the interrelationship of haplogroups and SNP test results.[9] To explore the possible application of such trees to our Study I am much indebted to Rick Byers who has produced simple trees using UPGMA and Kitsch software, and applying individual mutation rates to provide a conventional timescale. I asked him to develop these as a comparator for the work of Bill Howard who has applied Mathematica software with a timescale calibrated not from individual marker mutation rates but from correlating over 100 time/RCC relationships from a handful of separate pedigrees.[10]

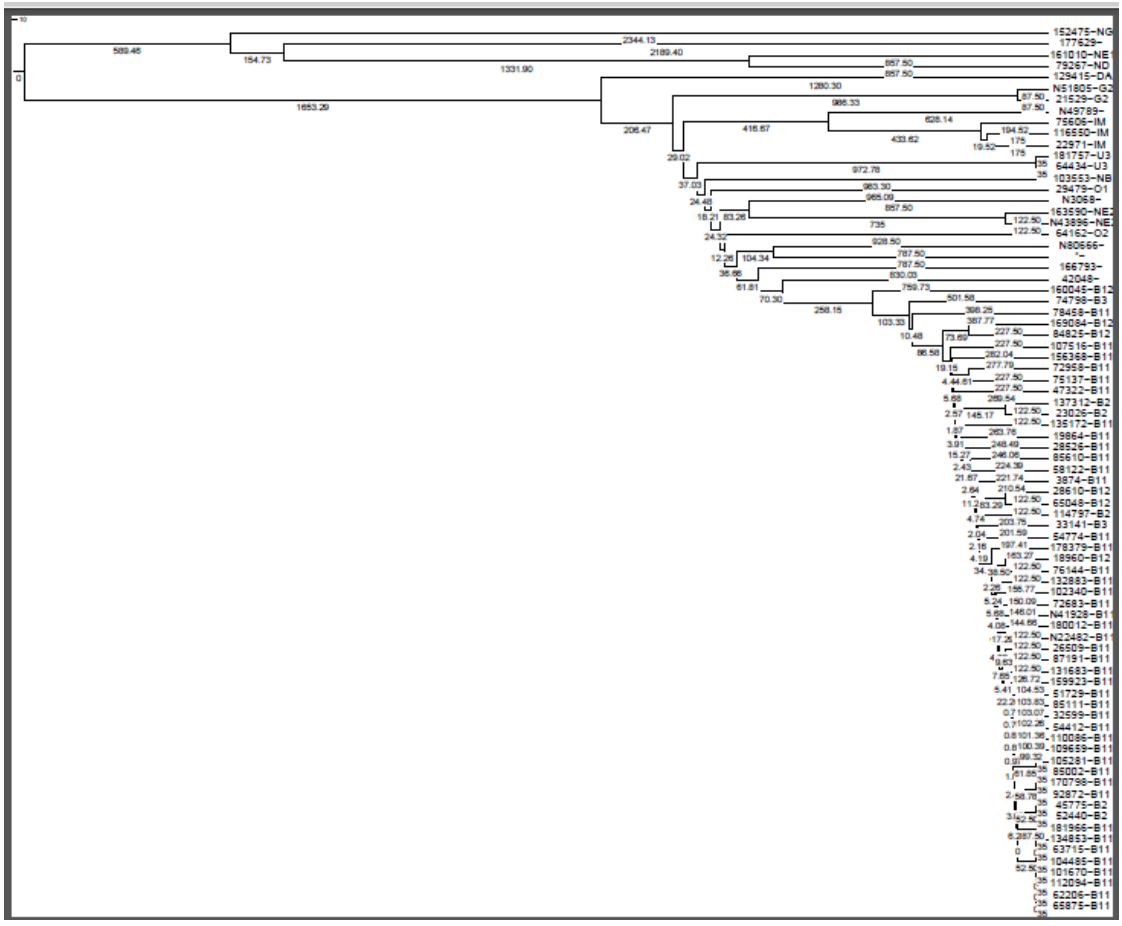
These phylogenetic trees can be compared at two levels:
(1) Using our full database, to check the genetic families previously identified in our Study; and
(2) Using our Borders database, to check the tentative sub-groupings thereof and identify new sub-groups.

5.1 Comparison of phylogenetic trees with full Irwin 37-marker database. The results are illustrated overleaf.

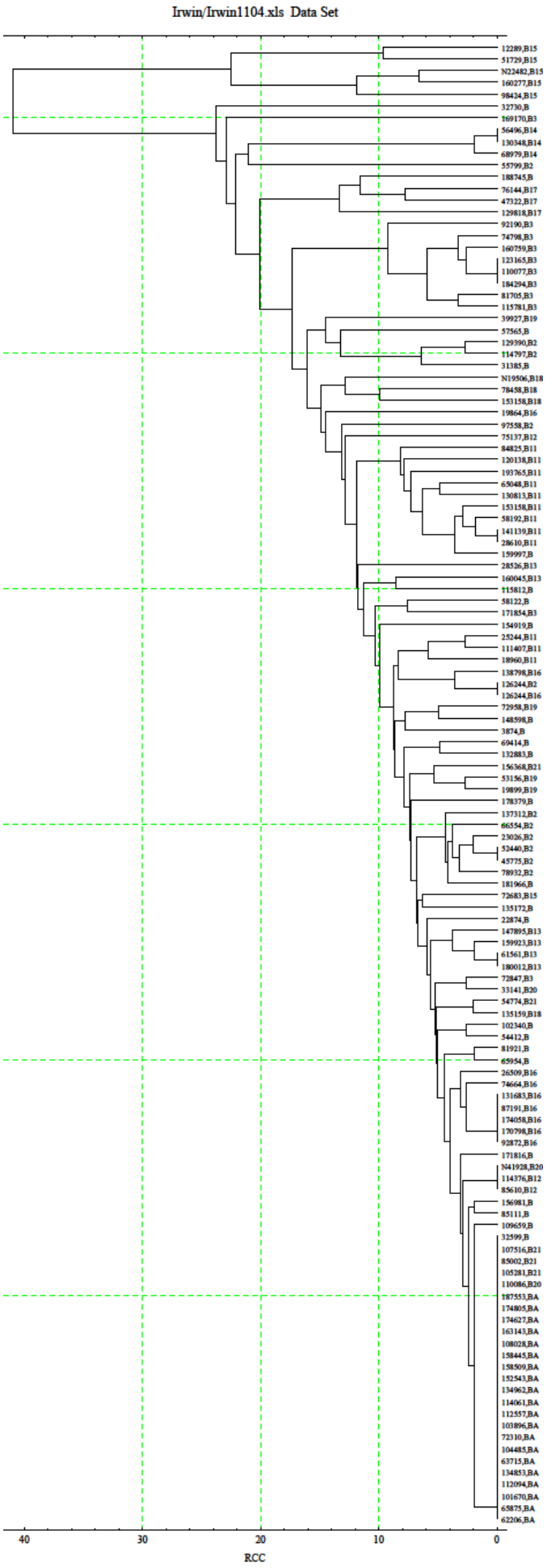
[9] The portion of FTDNA's phylogenetic tree of their new Haplogroups that is relevant to an individual participant is available on their personal page at Y-DNA > HAPLOGROUP.
[10] See http://www.jogg.info/52/index.html , http://mysite.verizon.net/weh8/Gordon20.pdf and http://mysite.verizon.net/weh8/Schwab.pdf



UPGMA phylogenetic tree



Kitsch pylogenetic tree



Mathematica phylogenetic tree

In each of these printouts the individual participants and their categorisation into genetic families as of February 2011 are listed on the right, ranked in order of bifurcation with the earliest bifurcations at the top.[11] The horizontal scale is time, with the present on the right, dating back to the left. For the UPGMA and Kitsch trees the evolved times are shown for each bifurcation; for the Mathematica tree the timescale is in RCC, where 1 RCC = c.50 years. But for all three the timescales are very approximate, their variability being much more than the trees imply.

Although these three reproductions do not do justice to the efforts of the individuals who kindly prepared them for me, and their presentation formats are not as refined as they or I would like, some clear conclusions can be derived:

1. Like network diagrams, phylogenetic trees can provide a clear visual depiction of how yDNA STR test results can be presented.

2. For divisions into genetic families, the three trees all show in solid blocks the genetic families that had already been identified by the TiP ranking procedure.[12] This consistency clearly confirms the validity of this procedure.

3. Philogenetic trees are very complex to prepare, and the application and interpretation of different software gives results with significant differences in their details. These differences are illustrated in the following summary analysis:[13]

[11] Both Byers's philogenic trees are shown jumbled and ladderized, and weighted using individual marker mutation rates as per Macdonald (this apparently had affect on sequencing, but extended lapsed time slightly). His TMRCA calculations are from http://www.mymcgee.com/tools/yutility.html and http://nitro.biosci.arizona.edu/zdownload/papers/MRCA.pdf using 50% probabilities.

[12] There are two minor exceptions: the Kirsch software separated one O2 participant (Kit 64162) from the other two O2 participants, and the Mathematica software excluded the NPE – Bell genetic family because these participants have a null count for DYS439.

[13] The non R1b1 singletons have been retained to help illustrate the underlying principles.

Haplo-group	Group/Cluster/Genetic family (R1b singletons excepted)	Code	No. of partic-ipants	Rick Byers **UPGMA** Sequence	years	**Kitsch** Sequence	years	Bill Howard C **Mathematica** Sequence	years
G	Singleton	SG	1	1	4415	1	4696	4	
I1	Singleton	SI1	1	5		3		5	
I1	NPE - Ellliot (1)	NE1	2	8		6		6	
I1	NPE - Kerr	Nke/r	1	7		5		7	
I1	NPE - Dodd	ND	2	6		4		8	
I2b	Ireland - Leinster	I?L	4	4		8		2	
J1	NPE - Graham	NG	1	2		7		3	
J2	Singleton	SJ2	1	3		2		1	14594
R1a	Undetermined - Drumkarney	UD	1	9		9		9	
R1b	NPE - Bell	NB	7	20		19		-	
R1b	Orkney (2)	O2	3	18		14		10	
R1b	Undetermined - Newton	UN	1	10	2948	10	2945	11	7217
R1b	Germany/Netherlands	G2	6	13		12		12	
R1b	Undetermined - B	UB/4	2	23		23		13	
R1b	NPE - Johnson	NJ	2	14		18		14	
R1b	Drum/Aberdeenshire	DA	2	15		11		15	
R1b	Perthshire - Fortingall	PF	2	17		17		16	
R1b	NPE - Kincaid	Nki/d	1	11		13		17	
R1b	NPE - Rutledge	NR	3	19		20		18	
R1b	Undetermined - A	UA/3	2	12		16		19	
R1b	Orkney (1)	O1	3	22		22		20	
R1b	NPE - Elliot (2)	NE2	2	21		21		21	
R1b	Ireland - Munster	IM	3	16		15		22	
R1b	Borders	B	125	24	1590	24	1380	23	1924
	" (earliest internal bifurcation)				1035		645		1775
	" (latest internal bifurcation)				50		50		120

This summary shows:

1. Although all the three trees rank the Borders genetic family participants below (i.e. later than) our other genetic families, they show significant differences in how they sequence these main bifurcations.[14]
2. The three trees have significant differences in the time elapsed since the bifurcations of the individual families. However these time differences relate primarily to “deep ancestry”, so their explanation and implications are outside the scope of our Study.

5.2 **Comparison of phylogenetic trees and network diagrams with the Border Irwins database.** At this level it is appropriate to re-introduce the Fluxus network diagrams developed by Rick Byers, and to compare two diagrams he has prepared with data extracted from the three phylogenetic trees.

The comparison table overleaf shows how these five software packages have grouped individual participants, designated by Kit No., who have already been identified as being within the Borders genetic family, and places them within existing and some possible new sub-groupings.[15] For the two network diagrams the participants are simply listed in order of their radial distance from the hub, which includes all the participants with the modal marker counts characterised as ‘BA’. For each of the three phylogenetic trees many of the participants in the existing sub-groupings B3 ,B2 and B12 appear consecutively in a block, as shown in the table, or elsewhere, either as mini-blocks or as isolated entries. The data for the three trees is presented first in (A) for the existing sub-groupings and second in (B) for some possible ‘improved sub-groupings’.

[14] Some of these differences may be more cosmetic than real due to ladderization of data.
[15] The existing sub-groupings BA, B12, B2 and B3 have been based on subjective criteria of prominent combinations of marker counts.

(A) With existing sub-groupings

	RByers28/3 Fluxus		R Byers 8/4 Fluxus			Rick Byers UPGMA		Rick Byers Kitsch		Bill Howard D Mathematica	
	No	Kit Nos	No	Kit Nos		No	Kit Nos	No	Kit Nos	No	Kit Nos
B3	12	72847	9	72847	main	8	92190	8	92190	8	115781
		33141		160759	block		115781		115781		81705
DYS		171854		115781			81705		81705		74798
449		160759		169170			160759		160759		184294
= 31		115781		?188745			74798		74798		123165
		81705		110077			110077		184294		110077
		123165		(123165			123165		110077		160759
		92190		81705			184294		123165		92190
		74795		92190							
		(110077			isolated		171854		171854		161970
		169170			(*: mini-		169170		169170		171854
		(184294			group)		33141		33141		*33141
							72847		72847		*72847
B2	11	52440	7	52440	main	4	23026	3	137312	5	66554
		78932		(45775	block		78932		23026		78932
DYS		23026		97558			45775		78932		23026
449		137312		78932			52440				45775
= 29		126244		181966							52440
		45775		?114797	isolated		55799		55799		55799
		97558		?129390	(*: mini-		97558		97558		*129390
		114797			group)		*114797		*114797		*114797
		129390					*129390		*129390		97558
		55799					66554		66554		126244
		66554					126244		126244		137312
							137312		*45775		
									*52440		
B12	14	18960	11	28610	main	13	84825	6	130813	9	84825
		111407		(141139	block		193765		65048		193765
DYS		138798		?18960			120138		169084		65048
391		65048		111407			130813		58192		130813
= 10		130813		(?			65048		141139		169084
		28610		84825			169084		28610		28610
		(141139		?193765			58192				141139
		58192		130813			141139				58192
		160045		126244			28610				120163
		84825		120163			160045				
		?120138/63		?129390			111407				
		25244					25244				
		169084					(138798				
		193765									
					isolated		18960		*120138		160045
					(*: mini-				*193765		138798
					group)				*84825		*18960
									160045		*111407
									25244		(25244
									*111407		
									*18960		
									(138798		

(B) With sub-groupings revised

		Rick Byers UPGMA		Rick Byers Kitsch		Bill Howard D Mathematica	
		No	Kit Nos	No	Kit Nos	No	Kit Nos
B3	main	9	39927	8	92190	9	39927
	block		92190		115781		115781
(DYS			115781		81705		81705
449			81705		160759		74798
= 31			169759		74798		184294
and			74798		184294		123165
385b			110077		110077		110077
= 14)			123166		123165		160759
			184294				92190
	isolated	1	154919	2	39927	1	154919
					154919		
B12	main	9	84825	6	130813	9	84825
	block		193765		65048		193765
(DYS			120138		169084		65048
391			130813		58192		130813
= 10			65048		141139		169084
and			169084		28610		28610
CDYb			58192				141139
= 39			141139				58192
or 40)			28610				120163
	isolated	0		0		0	

This complex and tedious comparison shows that all five software packages imply the three existing sub-groupings may not be optimal. Some possible improvements in the existing sub-groupings indicated by the three phylogenetic trees are shown in Table (B), but these improvements do not seem to be supported by the network diagrams, which do illustrate some inherent ambiguities in the data (where radials re-converge) and deficiencies (nodes not yet represented by data points).[16] And none of the five packages show any clear-cut advantage in identifying sub-groups (in marked contrast to their clear confirmation of our divisions into genetic families), let alone provide a stand-alone, objective tool by which these may be identified.

These rather negative findings are offset by two positive features:

- comparison of the network diagram data with the marker counts that can be associated with the sub-groupings already identified and partly confirmed by the trees does show that these sub-groups can be identified by a common count of a single marker, other than the CDY and DYS464 markers.[17]
- The Mathematica phylogenetic tree seems to offer some credible timescales.

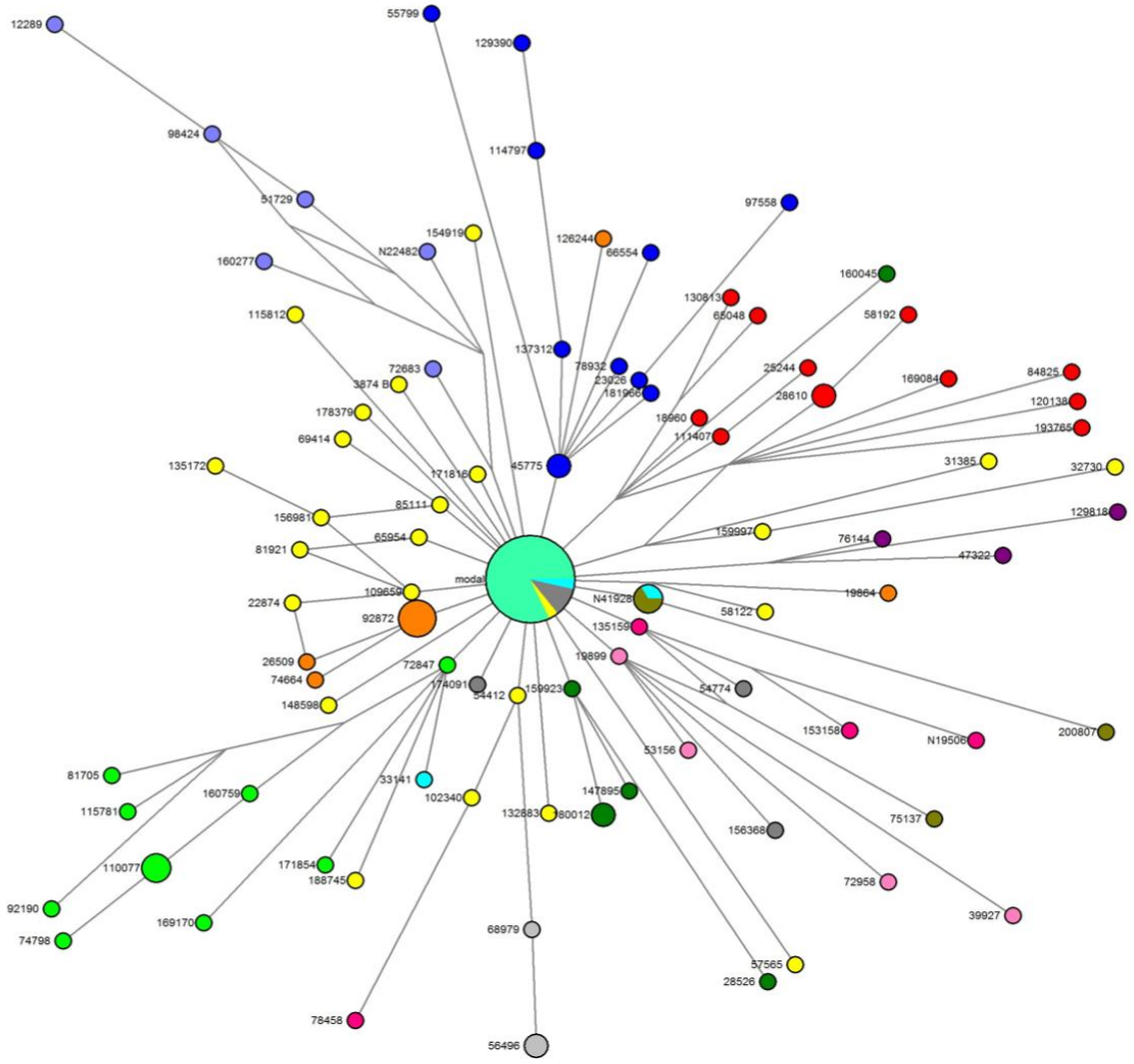
Procedure now adopted for identifying sub-groups.

From these observations on network diagrams and phylogenetic trees it is apparent that no one of the above tools can alone identifying and defining sub-groups. But given the lead that sub-groups can be identified by visual inspection for common counts of single markers, it became apparent that this "rule" is compatible with the limited amount of genealogical data we have, i.e. that it places the three Dumfries participants in a single sub-group and, after going to 67 markers, the three Bonshaw/Castle Irvine participants as well. There are however a considerable number of participants who qualify for more than one sub-group. For these, and for

[16] Such deficiencies are not surprising, given that less than 0.2% of Irwins have yet been tested.
[17] No similar associations with rare markers of combinations of markers stood out.

"ranking" participants within each sub-group, procedures developed for similar problems with genetic families[18] seem appropriate: identifying the "modal" participant within each group, comparing the 24-generation TiPs of marginal participants to see which sub-group they are more closely related to, and then ranking the participants of each sub-group by TiP comparisons with the modal participant. The absolute TiP values have no significance in this process. This provides a clear and simple procedure that has consistencies with that already adopted for genetic families.

The application of this process has enabled me to identify categorize 75% of participants in our Borders genetic family into 14 sub-groups of three or more participants that are compatible with virtually all the relevant genealogical data.[19] Having adopted this "rule" of categorizing sub-groups based on commonalities of single marker counts and, in cases of ambiguity, 24-generation TiPs, Rick Byers kindly re-ran a network diagram to test the rule against the "logic" of his application of the Fluxus model to the Borders 37-marker data. The sub-groups are identified with the prefix B followed, where possible, by a capital letter indicating their apparent place of origin (BE for Eskdale, BB for Bonshaw, BD for Dumfries)[20] or, for NPEs, two lower case letters (Bel for Elliot, Ber for Errand). The result is thus:



Network diagram of Border Irwins 37-marker data, 30 April 2011[21]

Key:							
	BA:	blue-green	Ber	light grey	B16	brown-green
	BB	dark grey	B9	dark green	B17	hot pink
	BD	orange	B10	red	B23	light purple
	BE	light green	B14	dark purple	B29	dark blue
	Bel	pink	B15	light blue	B	yellow

[18] For a detailed justification of this criteria, see Appendix C of the accompanying Supplementary Paper No.1 ("Towards Improving ...."). In brief: if the DNA signature of a single participant has a TiP of less than 80% from the modal value of the identified genetic families they are deemed to be a "singleton"; if two or more singletons have mutual TiPs of over 80% they form a new genetic family.

[19] With one exception: though 72683 and 85111 have been identified as 5th cousins, the former appears in B23, the latter is unmatched.

[20] Consequentially the indicators for the individual participants who enabled these identifications have become Be, Bb and Bd etc.

[21] Multi-node identifiers: "Modal" also includes 105281*, 85002*, 110086*; "45775" also includes 52440; "28610" also includes 141139; "N41928" also includes 85610*, 114376*; "80012" also includes 61561; "56496" also includes 130348; "110077" also includes 123165, "92872" also includes 170798, 174058, 87191, 131683. NB: The kit nos. with "*" are those in the mixed-group nodes, which occur because the BB and B15 sub-groups depend on 67 markers, whereas this cladogram is restricted to 37-marker data.

Given that the sub-groups BB and B15 are based on 67-marker commonalities and B are unclassified, this suggests a good correlation for all the proposed sub-groups except B16. Some individual exceptions, for example Kit Nos. 160045 and 126244, are instances where there are ambiguities in the marker count commonalities, and some detailed refinement may be appropriate.

Having identified the sub-groups the next challenge was to have the Mathematica phylogenetic tree software run again, using identifiers for these sub-groups to see if the sub-groupings could be dated.[22] Unlike with the genetic families, the sub-groups did not all appear as solid blocks on the tree, but for most sub-groups there were main blocks whose mutation dates could be dated relative to the rest of the family.[23]

The results of all this work are listed in columns (1) through (5) in the table below, sub-groups being sequenced in date order, starting with the youngest. Some sub-groups cannot yet be dated: the small B16 and Bel sub-groups because they do not appear as solid blocks; and the BB and B15 because their identifications are dependent on common markers which occur in the 38-67 panel, for which Bill Howard's Mathematica package is not yet calibrated.

Sub-group		Partic-ipants	Mutation date			Earliest genealogy	Mutation location	Provisional categorization
	(1)	(2)	RCC (3)	years ago (4)	AD (5)	(6)	(7)	(8)
BA		23	3	150	1800	1700	Scotland	pre 1700
B15		3	?			1700	?	pre 1700
BD	Dumfries	10	4	200	1750	1600	Scotland	pre 1600
B16		4	?			1800	?	pre 1800
Bel	Elliot	4	?			1800	Scot. or Ireland	pre 1800
B9		6	5	250	1700	1700	?	pre 1700
B29		11	5	250	1700	1650	?	pre 1650
BB	Bonshaw	6	?			1500	Scotland	pre 1500
B10		12	12	650	1300	1700	?	"old"
B17		4	15	750	1200	1750	?	"old"
BE	Eskdale	11	17	900	1050	1600	Scotland	"old"
Ber	Errand	3	22	1150	800	1850	?	"pre-surname"
B14		3	23	1200	750	1750	?	"
B23		6	41	2150	BC200	1650	?	"
B	unmatched	37						
Total		143						

Before discussing this data several points have to be emphasised. First, the categorization is not rigid:

- some participants "qualify" for more than one sub-group;
- double mutations of a single marker of an ancestral line may invalidate its sub-grouping;
- my rules for sub-grouping may not be optimal, and the BE sub-group may be misidentified.

Second, despite the apparent precision of the mutation dates for each sub-group in the table above, they are subject to a number of errors. Some of these errors are due to the randomness of the mutations that underlie the whole subject; some, probably of a similar order of magnitude, are inherent in the derivation and calibration of the RCC time scale. The errors are greater when the resolution of the markers is small, when sample size is small and, in percentage terms, when the mutation occurred in relatively recent times. In the context of this Study the sum of these errors probably have a standard deviation in the order of 300 years, i.e. there is about a one-in-three chance that any of the mutation dates above may be in error by more than 300 years either way.[24] It is also important to note that the mutation dates are not absolute but are relative to the rest of the family, and where the sub-groupings are "fuzzy" the associated datings become ambiguous. It seems likely that as more participants join the Study and the sizes of these sub-grouping grow then the clarity of these datings should improve.

The sub-groups may also be dated from genealogical ("paper trail") data: for those participants in each sub-group who are not apparently genealogically related, the mutation characterising their sub-group must have occurred before the earliest confirmed paternal ancestors of such participants. In other words, both the approximate mutation date (column (5) in the table above) and the date of the earliest unrelated paternal

[22] Although the UGMA and Kitsch diagrams show timescales, no effort has been made to optimize the many input parameters that affect their magnitude. They are included to illustrate their potential as comparators rather than to develop meaningful timescale forecasts.
[23] I am indebted to Bill Howard for enabling the identification of these dates from the phylogenetic tree he prepared from our data, and for his patience in explaining to me how this data can be interpreted.
[24] This 300 year error is endemic to all approaches and is largely due to mutations. The Mathematica model reduces the error when it performs optimization (advice by Bill Howard). The context of this 300 year error is considered in a footnote to this paper.

ancestor (column (6) in the table above) represent the latest dates, genetic and genealogical respectively, when the bifurcation of the sub-group from the rest of the genetic family could have occurred.[25]

Taking all these considerations into account, the conclusions that can be drawn on the apparent ages of our 14 sub-groups (column (8) in the table above) are much more limited than the probable dates that the Mathematica model implies. We are probably only justified in noting that:

- None of the sub-groups are "younger" than c.1800, and all may predate migrations from the Borders to Ireland during the 17th century.
- For the four sub-groups (BA, BD, B9 and B29) there is reasonable correlation between the earliest genetic date and the earliest genealogical date.
- The first seven sub-groups (BA, B15, BD, B16, Bel, B9 and B29) are probably relatively "young".
- The next four sub-groups (BB, B10, B17 and BE) are probably relatively "old", and probably date back to between c.1500 and the era when the surname was first used in the Borders (i.e. the period between 1376, the earliest surviving record of the surname there, and c.1250, when hereditary surnames were probably first introduced in southern Scotland). This is consistent with surviving records of Irvings residing in the Dumfriesshire homesteads of Bonshaw, Gretna, Hoddam, Luce, Pennersax, Skaill, Stakeheugh and Turnshaw between 1400 and 1500.
- The last three sub-groups (Ber, B14 and B23) appear to pre-date the surname era. This suggests that for the Ber sub-group, Errand is not a misspelling of Erwin as had been thought possible, but was always a quite different surname, even though it has a similar DNA signature, and its members may thus be descendants of a NPE like the Armstrongs in sub-group BA. The much older B23 sub-group is even more distantly related genetically, and an ancestor of its members may have adopted the surname by co-incidence, or simply as neighbours. The B14 sub-group may be like B23, or a misdated "old sub-group".
- One implication these findings is that although thee modal sub-group BA is the largest sub-group, its DNA signature probably does not represent that of the common ancestor of the Borders genetic family. At present it is unclear which of the four or five "old" sub-groups is the most likely to represent this common ancestor, if indeed his DNA signature still survives today.

Perhaps a more important implication than these very tentative thoughts on dating is that the identification of these 14 sub-group now enables most individual participants within the Borders genetic family to identify themselves with a sub-group that includes other participants to whom they are likely to be genealogically related. But it has to be remembered that likelihood of participants being able to identify such relationships has three important limitations:

- the sub-groupings now developed, while objective and rational, are predictive rather than determinant;
- as we have seen, random mutations can make individual marker counts unreliable, and hence consequential assignment to a particular sub-group;
- available genealogical evidence may be unreliable or too recent to confirm the sub-groupings.

The extent of these limitations will determine the success rate when individual participants within each sub-group attempt to confirm the genealogical relationships that these sub-groupings predict.

[25] There is a third way in which we can tentatively date our sub-groups. If a sub-group includes participants whose ancestors never left Scotland and also participants who are now resident in USA (or elsewhere in the new world), then it would follow that the mutation that characterizes this sub-group must have occurred in Scotland, before any ancestral migration (assuming no reverse mutation has subsequently occurred). And given the probability that ancestors of many of our American participants began their migration from Scotland, via Ireland, to USA in the 17th century, it would follow that the mutation date characterising sub-groups with participants living in both Scotland and USA was probably before c.1700. Alas this logic has two shortcomings: firstly, it does not apply to the opposite contingency, i.e. for sub-groups that include no "old world" participants we cannot deduce that the relevant mutation occurred after migration, because it may be that there are some old world members of the sub-group alive to day who have yet not undergone a DNA test; and secondly, it so happens that at present this logic does not help to reduce the dating uncertainties for any of our individual sub-groups. On the plus side, however, this logic is not inconsistent with earliest confirmed ancestor paper trail data – and hopefully with more data these tentative dating estimates will become more precise.

Footnote

Having recognized there are significant limitations to dating sub-groups using this application of the Mathematica model it is salutary to consider the limitations of the "classic" TMRCA (Time since Most Recent Common Ancestor) models that depend on the genetic distance between two individual testees. Thus for example at http://dna-project.clan-donald-usa.org/tmrca.htm , the following data can be derived: [26]

Genetic Distance	Most probable TMRCA			2/3 of TMRCAs within	95% of TMRCAs within
	transmission events	generations	time elapsed		
0/37	1	$^1/_2$	15 years	30 - 135 years	0 - 225 years
1/37	8	4	120 years	100 - 400 years	20 - 700 years
2/37	16	8	240 years	200 - 550 years	30 - 900 years
3/37	26	13	390 years	300 - 700 years	60 -1100 years
4/37	34	17	510 years	350 - 900 years	110 -1300 years
5/37	44	22	660 years	500 -1100 years	170 -1500 years

Sighting the 37-marker genetic distances from BA for participants in the subgroups of the Borders genetic family (column (16) of the Results table) shows how inadequate genetic distances and TMRCA data are for estimating the age of a small sub-group. The sub-group with the least "spread" of genetic distances is BE, with genetic distances from BA of 3/37-5/37, implying the mutation bifurcation of BA from BE most probably between about 390 and 660 years ago, but with 95% confidence we can only say between 60 and 1500 years ago – in practice an almost useless finding. Similarly for the Bel sub-group, whose four members have genetic distances of 1/, 2/, 3/ and 4/37 respectively, the 95% TRMCA is between 20 and 1300 years ago!

[26] The times elapsed assume 30 years per generation. All the above assume a default mutation rate of 0.0033 per transmission event. If instead (arbitrary) mutation rates of 0.0020 and 0.0045 are assumed, the most probable TMRCAs for a genetic distance of 1/37 would become 90-210 years respectively; for a genetic distance of 5/37 they would become 450-1000 years respectively. For 67 markers, assuming the 0.0033 rate, the most probable TRMCAs for 1/67 and 5/67 are 75 and 180 years respectively.