

Colla and its Wider Diaspora

(How the Colla population fits in to the larger R-L21 population)

Patrick McMahon ©
Mount Alexander
Gorey
Co. Wexford
Ireland
(*mcmahongorey@utvinternet.com*)
January 2012

The views expressed in this paper are largely those of the author. The author would like to acknowledge and thank his co-administrators (of the Colla Project), Josiah McGuire, Peter Biggins and Thomas Roderick for their contributions during the course of this study.

Summary

This evaluation was conducted in order to bring together the various observations made over the last few years about Clan Colla and particular DNA profiles. In the early days of the Colla Project, an association was made between a specific mutation (a null mutation¹ at DYS 425) and typical clan names known to be from Oriel (County Monaghan and parts of surrounding counties)². It was quickly established from the limited records (Irish annals and the like) that these Oriel names were claimed to be descended from the 'Three Collas', alleged mercenaries brought to Ulster from Britain in the 4th century at the behest of the Irish High King. The overall aim of this work was to establish the credence of this history and the emerging DNA story.

Implicit in the history is that the Three Collas were of British Celtic descent. The origin of these British Celts was thought to have been north of the Alps, whose defining mutation was R-L21+ thought to have occurred about 4,000 years ago. The archaeological evidence and observed spread of L21+ supports the view of a North Westward drift of this population through Germany, France, Britain and Ireland, culminating in it being most concentrated in Ireland and the Western fringes of Britain. This analysis has placed the Colla population firmly within this much larger R-L21 Celtic population and reflects some of the later migratory moves. Emerging data has further refined the Collas as being within the R-DF21 sub-clade of R-L21. The observed results in the evolutionary tree are consistent with the spread of R-L21 throughout Europe, Britain and Ireland. The Colla population currently constitutes about 16% of the R-L21+ population.

Using statistical methods, it has been possible to construct an evolutionary tree showing the interrelationships of deduced Colla ancestors and other non-Colla ancestors within the R-L21+ population. There is clear evidence for the point start (acquisition of the null mutation) of the Colla population within R-L21+ consistent with the historical records. It is also clearly evident that many Oriel names have also got non-Colla DNA lineages. As a general observation it can be stated that lineages which had branched early in the evolutionary tree tended to be lone representatives whereas those branching more recently tended to be part of clusters. The degree of relatedness in general was proportional to the time of branching.

The appearance of so many 'non-Celtic' names among Colla DNA descendants was at first puzzling and attributed to adoption and other name changing events. Although some are known adopted names and more are thought to be, the majority are probably not adopted names. By far and away the easiest explanation is to consider these deduced ancestors as being siblings, one evolving in situ (say Scotland) and the other migrating to Wales or Ireland and evolving there. Many of these migrations occurred before the onset of naming and so the respective devolved lineages would have acquired the name appropriate to their location and clan association at the time of naming (about a thousand years later). This explains how the evolutionary tree can show a typical Oriel name linked to a typical Welsh or Scottish name inferring that one name changed into another which was not the case.

¹ A Glossary of terms used is at Annex F.

² This is attributed to the Clan Colla Project Administrator, Josiah McGuire.

1 Introduction

The understanding today is that Clan Colla descendants are R-L21+ people who carry a null value for DYS marker 425 in their DNA, have generally a Celtic name and a genetic distance (GD) of no greater than 11 to the Colla modal (DURRQ). This definition has served us well over the last few years and has proved to be a robust one. Without doubt there is a Colla population whose numbers now exceed 200 (testers) who meet the above criteria. Lately however, a number of anomalies have appeared in increasing numbers which are not easy to explain. The main anomaly is how to explain the very high proportions of Colla Celtic names (such as Carroll, McMahan, McGuire) who do not meet the above criteria and equally, the significant number of non-Celtic names who do. The present analysis was initiated in order to try and answer these anomalies.

2 Present Understanding

Recent work established that there were seventeen clusters among the 160 profiles examined and that the precursors of these had diverged prior to the allocation of surnames. Early branching suggested the approximate date of birth of the Colla Progenitor to be c. A.D. 280 (Ref 1). The Colla Progenitor gave rise to a population of Collas, probably in NW Britain, who expanded into Dal Riada and Ulster. Most but not necessarily all of the 'Irish' Collas are allegedly descended from three Colla brothers who came to Ireland in about A.D. 340.

Early divergence is the most likely explanation as to why genetically diverse lines (who were part of the same tribe), finished up with the same surnames. Examples of this are the two McGuire Groups and the three McDonald Groups as well as several singletons. Equally, it goes some way to explaining mixed origins where $\frac{2}{3}$ to $\frac{3}{4}$ of many of the major Colla names did not have Colla DNA. The precursors of these non-Colla DNA McMahons, McGuires, Carrolls etc must have been already associated³ with the Colla tribes prior to the allocation of surnames (Ref 2).

Given the wide variety of names, many of which are 'non Irish'⁴, it is hard to escape the conclusion that the origin of the Colla clan was among the British Celts. If as alleged, the Colla Brothers arrival as Roman trained mercenaries was at the behest of the High King of Tara, then this would support this view. Genetically, the closeness of Calkin to McMahan, Carrolls and many others is strong supportive evidence. Similar arguments can be extended to Rodericks, McDonalds, and Paden. The simplest explanation is that however many sons Papa Colla had, they or their progeny spread over North West Britain, Dal Riada and Ulster, their descendants all showing relatedness to this day. Around A.D. 340, three of his sons came to Ulster (leaving relatives behind presumably) as mercenaries from whom most, but not necessarily all, of the 'Irish' Collas are descended. For example, somebody living in Dal Riada (prior to the allocation of surnames), whose relatives would eventually be called McDonald, might have moved to Fermanagh where his descendants might be called McGuire (44801 McGuire in this example). Genetically, their descendants would show up in a cluster, as in this case in a McDonald cluster.

³ Some tribesmen who lived with or near the Colla clan could have taken on the names McMahan, McGuire, Carroll etc. Even with a strong oral history tradition and over a period of 20 generations, it must have been difficult to keep track of who were pure male Colla descendants especially as the Brehon Laws encouraged concubinage.

⁴ Names such as Calkin, Roderick, Paden, McClain, McDougall, Walden, Dever, McDonald, Rattray, Rice, Godwin, Kern, Judd.

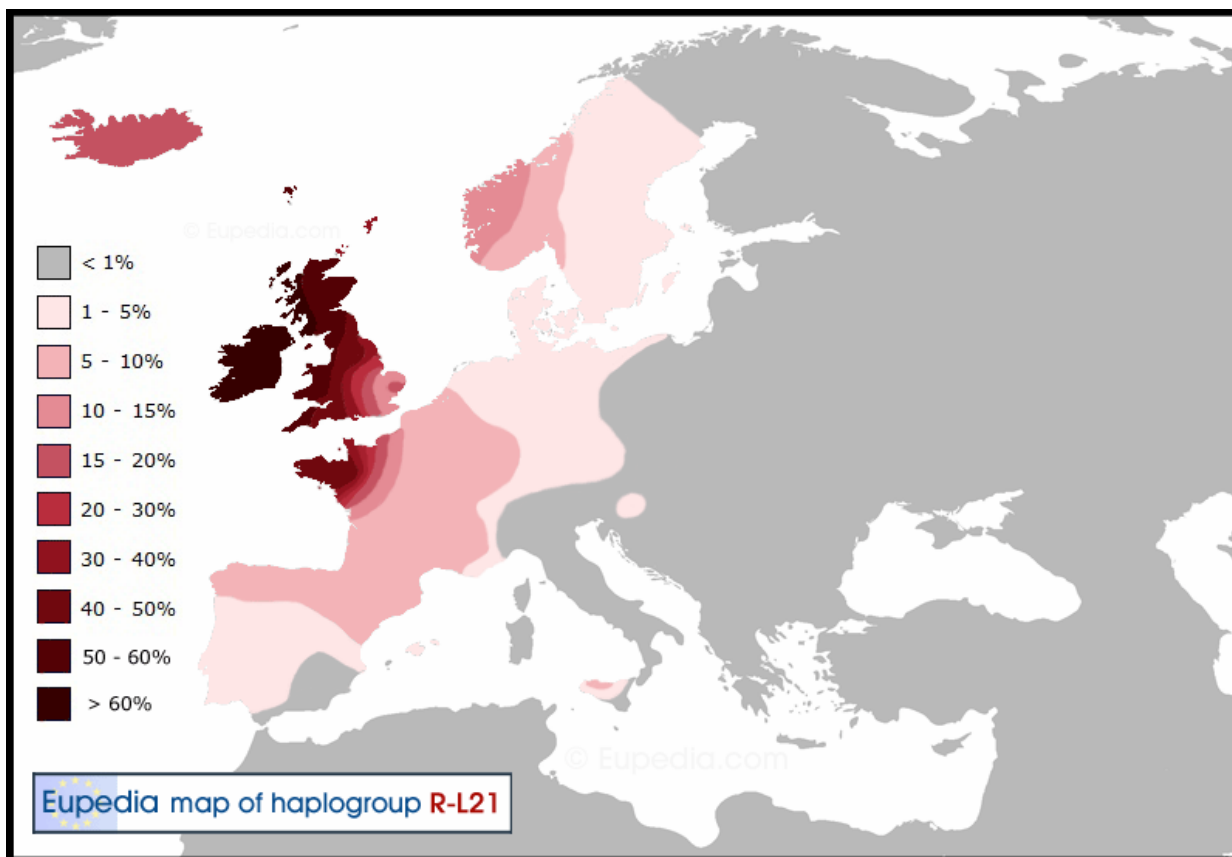


Figure 1. Distribution of R-L21.

It is thought that a population of R-L21+ people arose (by acquiring the L21+ mutation) north of the Alps about 4,100 years ago. R-L21+ descendents spread through France, the Low Countries, Britain and Ireland (Figure 1, Ref 3). It is estimated they could have started to populate Ireland shortly after this. By the time the Romans came to Britain, these Celtic people would have subsumed earlier cultures and diverged from one another over the 50 or so generations. They developed into tribal groups as they did in Ireland which the Romans, for their administrative convenience, gave them names such as the Brigantes, Trinovantes, Cornovii, Silures etc. The L21+ mutation is most concentrated in Ireland (60-70% according to the Co-administrator of the "Ireland yDNA" Project) and to lesser extents in parts of Britain and north-western Europe. This supports the findings in Figure 1 showing a north-western migration from continental Europe over time. It is thought that these British Celtic tribes⁵ were mainly R-L21+.

3 Aim

The aim of the current work was to explain the relationships between members of the defined Colla family (as given above) and those sharing a similar genetic makeup (L21+) and to see how they relate to each other. These distant relatives included those having nulls and no nulls at DYS 425 along with specific values for markers 511 and 413.

4 Assumptions

The Colla administrators are agreed that all Colla descendants are R-L21+ in so far as all of those who deep clade tested proved to be R-L21+ and none proved to be R-L21-. It is therefore a reasonable

⁵ The [Bell Beaker Culture brought the Bronze Age to the British Isles](#). To be more exact, Beaker folk initially brought the Copper Age around 2,450 BC, homing in on the copper belts of Ireland and Wales. They left their characteristic beakers at a copper-mine on [Ross Island](#), in Lough Leane, County Kerry. To judge by chemical composition, copper from Ireland was traded into Britain, along with gold from the Mourne Mountains. The incomers boosted what had been a dwindling population of farmers, and created a thriving society. From around 2,200 BC Bell Beaker interest in Britain intensified as Cornwall was discovered to be a prime source for tin, the rare and precious component of true bronze. This resource gave the British Isles a head start in Europe in making bronze.

assumption that Collas are a sub-set of R-L21+ and that genetic relatives will be found among R-L21+ people.

The remaining assumptions for this study were:

Virtually all Collas had the following values for three key markers 511=9, 425=0, 413=22-23. Values for these markers were used to track closest relatives.

That the closest genetic relatives would be found with variant values for these three key markers.

That the Kitsch programme (one of many in the PHYLIP collection) using the Fitch & Margoliash methodology was able to differentiate the relative branching events in the dataset from the aggregated marker values. The implicit assumption is that this will only apply to members of the same population. Where two populations (i.e. Colla and non-Colla) are examined in the same batch, the relative branching positions of the non-Colla population may not be accurate as it could have descended from other than the Colla precursor.

5 Limitations

The limitations were associated with the statistical methods used in McGee (Ref 4) to calculate tMRCA (time to most recent common ancestor) values which overestimated the number of generations for low GD values. The limitation in the PHYLIP methodology was underestimating the number of generations to the more distant branching (high GDs) (Ref 5). It was observed from previous work that cross relationships in McGee between those with low GDs and the rest of the testers is extensive. This is not unexpected as all are referenced to the modal making it easier for those with lower GDs to be related. Relationship is defined as a GD up to 6 (Green in McGee) irrespective of name. Table 1 shows the results obtained from 141 Collas having GDs ranging from 1 to 11.

For GD to DURRQ	No of Occurances	Average No of relatives
1	1	63
2	2	52
3	12	37
4	26	21
5	22	15
6	23	6
7	20	6
8	20	5
9	10	5
10	3	6
11	2	6

Relative defined as GD up to 6
(at 67 markers in McGee Utility)

Table 1. Distribution of GDs and relatedness among 141 Collas.

It is virtually impossible to visually identify molecular clusters (along the diagonal) for GDs up to 3. It gets easier for 4 and above and best for the highest GDs. This is why 628Hancock (GD 1), 514Devine (GD 2) and 189Main (GD 2) show significant relationship to so many (63, 62 & 41 respectively) and why Rodericks and McGuires have such 'clean' clusters.

6 Methods

This analysis is based on differences in the numbers of DYS mutations between individual profiles after conversion into tMRCA values in the McGee Utility. Some markers are more stable than others, some mutate concurrently with others and there are no measured rates for about half the markers. In addition, many of these markers can revert, reducing the observed GD. The only practical way forward is to use an average mutation rate. The range of average rates varies between .002 and .004

corresponding to 7.58 and 3.79 generations per mutational event. A rate of .00272 (McGee custom rate) was used in this analysis which equates to an average 5.7 generations per mutational event or 167 years/mutation (@ 30 yrs/Gen).

6.1 First Experiment

The sample of R-L21 non-Collas used was of necessity, limited. In the first experiment, testers with Colla and non-Colla names were selected in order to position the Colla sub group within the very much larger R-L21 population.

A specific database was constructed to test the above hypothesis which consisted of five well defined Colla clusters (McDougal, McGuire Gp2, Roderick, McDonald Gp1, Smith Gp1) and a selection of non null L21+ testers.

The entire R-L21 Project database (1101 data entries @ 67 markers, as at 22 Oct, 2011) was downloaded into Excel, sorted first by DYS 511, then by 425 and finally by 413. This ranked the data so that a selection of entries could be made starting with 9/0/21-23⁶ (36836 Peden) and finishing with 10/12/24-25. To these were added 3 McMahons who were 9/12 and 7 who were 10/12 together with 17 other proven L21+ who were 10/12 and had other Colla names. The five Colla clusters and a few Colla singletons were used as the reference Collas (which comprised 23% of all Project Collas). For ease of tracking, labels and colour coding were used as in Table 2.

DYS511	DYS425	DYS413	Label
9	0	21-23	
9	0	22-22	A
9	0	22-23	A
9	0	22-24	A
9	0	23-23	B
10	0	22-23	C
10	0	23-23	D
9	12	20-23	E
9	12	21-23	F
9	12	21-24	F
9	12	22-22	G
9	12	22-23	G
9	12	23-23	H
9	12	23-24	H
9	12	23-25	H
10	12	21-23	J
10	12	22-23	K
10	12	23-23	L
10	12	24-24	M
11	12	21-23	N
11	12	22-23	O

Table 2. Permutations of DYS 511,425 and 413 in the R-L21 Project. Label is the identifier used in PHYLIP and all tables.

6.2 Second Experiment

In the second experiment, individual members of clusters, which could be identified from McGee, were not examined; one member of each cluster was used to represent the lineage. The Collas were represented by 73 lineages (single member from each cluster plus singletons) and a sample of non-Collas (50) having a gradation of GD values (to DURRQ) beyond 11 in order to refine positioning of the Colla Group. To achieve the best resolution, 99 iterations were performed in the Kitsch Programme.

⁶ Values for the markers 511, 425, 413 are represented as 9/0/21. 413 is a multiple allele marker and where the second allele value is used, it is preceded by a hyphen, for example 21-23.

6.3 Name Groups and Clusters

For both experiments, the starting points were the name groups downloaded from the Clan Colla 425 Null Project. These consisted mainly of same name groups which were not always a good indicator of homogenous DNA⁷. Additionally, a few name groups included names which had Colla DNA but were known or suspected to have been adopted. Admitted adopted names were Hancock (GD of 1 to Devine), Walker (presently grouped with Padens), Stevens (born as a Calkin), Bogner (grouped with McMahon), O'Toole and Dundas (grouped with a Carroll).

In the second experiment, clusters were deduced in McGee by examining small numbers of name groups at a time (for clarity), bringing together those exhibiting GDs of 6 or less. A single representative from each cluster was used to represent that lineage, those having tested positive for DF21 being given preference. Testers who were not unambiguously part of any cluster were treated as singletons.

6.4 Excel

Excel was used to tabulate the data and organise it for inputting into the McGee Utility (Ref 4) and the generation of PHYLIP infiles. Each entry had an identifier made up of the first three digits of the FTDNA ID followed by a truncated⁸ surname.

Colla 9-0 (A)	137McDonaA	864McGuirA	49962RodeA	394McMahoA	740SmithA
	124McDonaA	170McGuirA	168RoderiA	138McMahoA	662SmithA
	112McDaniA	N35McGuirA	49959RodeA	431SandisA	135SmithA
	146McDaniA	108McGuirA	8551RoderA	393McClaiA	603SmithA
	133McDonaA	865McGuirA	8553RoderA		725SmithA
	688EdwardA	158McGuirA	49963RodeA	738KellyA	
			134RoderiA	111WaldenA	
			167RoderiA		
			124RoderiA		
			49960RodeA		
10-0 (C & D)	219McDougC	196McQuilD			
	144McDougC	929JohnsoD			
		E85MonterD			
9-12 (E,F,G,H)	319BushE	147StoneyF	N55BorcheG	258McMahaH	244BurdenH
	368BushE	112ByrneF	821DuncanG	718MahonH	959JonesH
	137McKenzeE	N55KearonF		N19MahanH	174SimonsH
	973LyonsE	128DohertF		851CoxH	123BroomH
	8042E	847PhilipF		134CorbitH	139021H
				160BodinH	173MorganH
				840HamonH	851MearshH
				112BasileH	865ThompsH
				168HorwilK	412MasseyH
				140003H	137DeMaschH
				167GibbonH	945PriceH
				179DavisH	N81SebillH
				N54ThomasH	138WilliaH
10-12 (J,K,L,M)	600CarroJ	112CarrolK	208McCaill	N59DuffyM	
	140CarroJ	401KelleyK	N65CarrolL		
	N42KellyJ		110McMahoL		
			113McMeekL		
			898McMaheL		
			142McMahoL		
			124McMahoL		
			143MaguirL		
			N20MaguirL		
			N30MaguirL		
			312McCownL		
			141McMahoL		
			150MahonL		
			107CarroL		
			N59CarroL		
			655DuffyL		
		312KelleyL			
		167KellyL			

Table 3. Composition of database used in Experiment 1 categorized as per Table 2

⁷ For example, having the McMahan surname only gave a 50/50 chance of being descended from the Collas.

⁸ The PHYLIP software had a limit of 10 digits per identifier.

6.4.1 Experiment 1

The composition of the database was as shown in Table 3. The identifiers had in addition one of the capital letters as per 'Label', Table 2. The database was examined in the McGee Utility Tool to identify relationships and generate the infile for PHYLIP. The settings used in McGee were Hybrid model, probability 95%, custom mutation rate, and generations.

6.4.2 Experiment 2

The data, labels, symbols and colour coding used are shown in Table A.1 (Annex A). Only the data associated with the listing under 'Lineage Rep' was used in the PHYLIP runs.

6.5 PHYLIP

PHYLIP (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). In using PHYLIP, the instructions put forward by Dave Hamm (Ref 5) have been broadly followed. Essentially, the programme (the Kitsch one was used throughout) compares all pair-wise combinations from the tMRCA matrix (generated in the McGee Utility) looking for nearest neighbours using the least squares method of Fitch-Margoliash (Annex C). For this analysis, the process used McGee tMRCA data (in generations) @ 95% p.

Where the ancestor of a population is defined by unique events, in this case the combination of 9/0 for markers 511 and 425 defining Colla, this method should produce creditable branching in the evolutionary tree. However, for non-Collas this might not always be the case. It would be expected that for those closest to the Colla ancestor (i.e. Colla precursors), this would still hold true but for those more distantly related (in terms of GD from DURRQ) there would be a chance that a particular profile could have arisen independently (in a different line) in the R-L21 population rather than be inherited through the Colla precursor line.

7 Results

The results are in three parts. The first part (sec 7.1) is a breakdown of all the entries in the L21 Project database (as at August 2011). Only those closest to the typical Colla profile of 9/0/22 (for 511, 425,413) were considered in this analysis. For marker 413 among the 9/0 L21+ category, 2% were 21, 92% were 22 and 6% were 23. The range considered was between 9/0/22 and 10/12/24. Profiles outside this range (termed 'Others' in Table 4) were counted but not analysed.

The second part of the results (sec 7.2) consisted of processing the 105 selected taxa (in Table 3) through the PHYLIP software. The resultant evolutionary tree branches were overlaid with the 511, 425, 413 values for each haplotype in order to track the mutational changes for these markers (Figure 2).

The third part (sec 7.3) was an examination of all the null carrying testers and their relationship to non-null kinsmen.

7.1 L21+ Distribution

The distribution of those examined (A to M) are shown in Table 4. Categories A & B are underestimated in the L21 database at 4.5%; by adding in the extra known Collas (from the Clan Colla 425 Null Project) brings this figure up to nearly 13% and could be as high as 16% or greater if all the unregistered Collas could be included.

As explained under Methods, a sample of 35 (23%) of known Collas was used as a reference in this analysis. Five out of seven 10/0, thirty-eight 9/12 plus three 9/12 Mahans were brought forward for analysis. By far the largest L21 category at 74 % was 10/12 but this category was restricted to 17 having Colla names in addition to 7 non null McMahons⁹. Those excluded from the area of interest amounted to 10% of L21.

⁹ Undoubtedly, many more from this category could be tested but the chosen sample was considered adequate to establish a 'proof of principle'.

Label	Grouped	No	%	Grouped	No	%	Additional*	%
A	9/0/22	46	4.18	}	9/0	49	154	12.78
B	9/0/23	3	0.27					
C	10/0/22	1	0.09	}	10/0	7	7	0.58
D	10/0/23	6	0.55					
E	9/12/20	5	0.45	}	9/12	38	38	3.15
F	9/12/21	5	0.45					
G	9/12/22	2	0.18					
H	9/12/23	26	2.36	}	10/12	887	887	73.61
J	10/12/21	116	10.55					
K	10/12/22	173	15.73					
L	10/12/23	585	53.18					
M	10/12/24	13	1.18					
	Others	119	10.82		119	10.82	119	9.88
	Totals	1100	99.99		1100	100	1205	100

* Plus Extra Collas not in L21 Project but in Colla Project.

If all Collas were registered (currently c. 200), their proportion would be 16% of L21.

Table 4. The distribution of haplotypes in the L21 database.

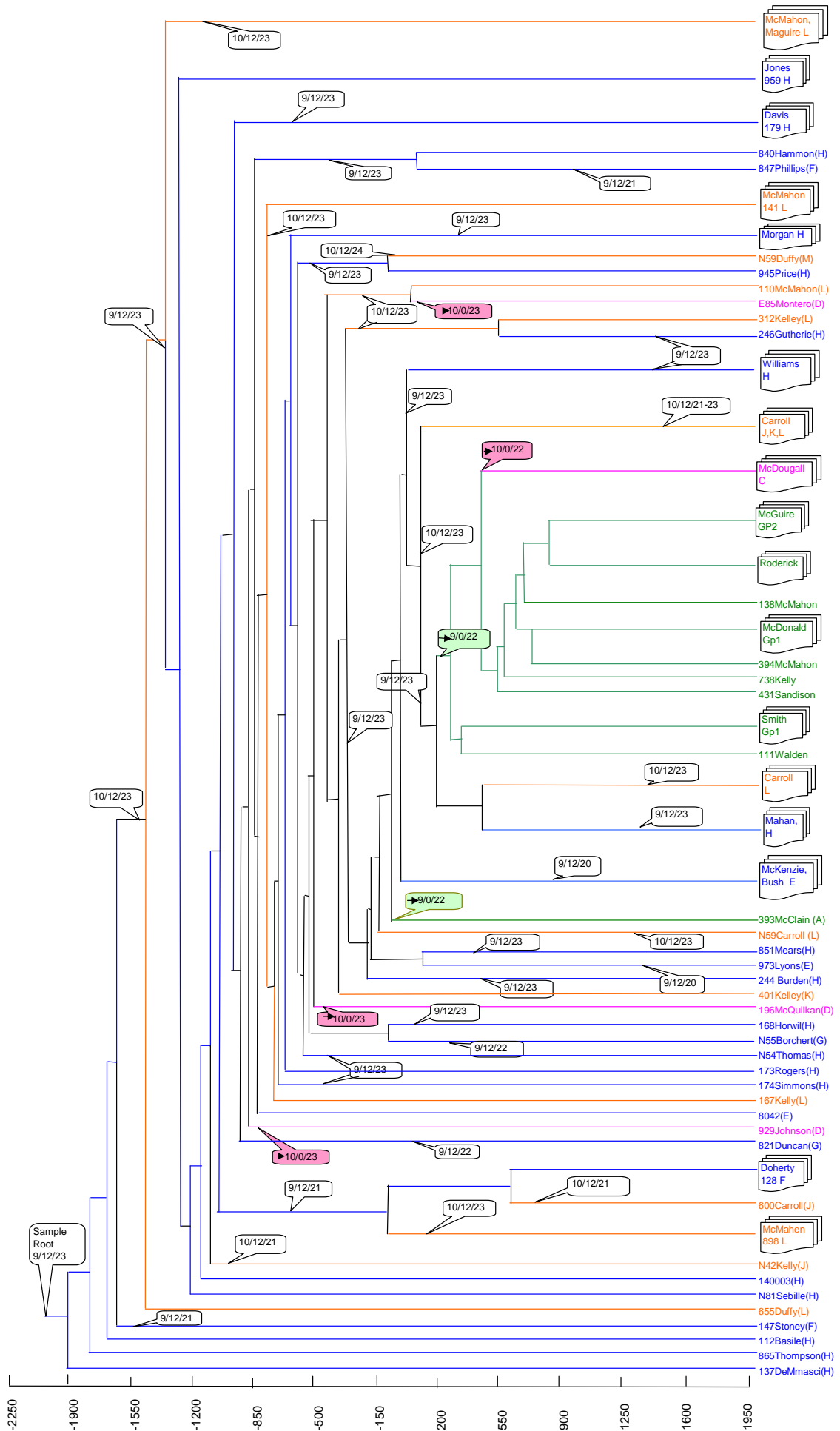


Figure 2. Schematic of the PHYLIP output in Figure B.1.

7.2 PHYLIP from First Experiment

PHYLIP output from a run with 11 iterations (jumbles) on the 105 taxa in Table 3 is shown in Figure B.1. For ease of viewing, this output was manually transcribed into the schematic in Figure 2. The schematic is somewhat simpler than the PHYLIP outfile, is colour coded as per Tables 2 & 3, and is overlaid with values for the three key markers on the evolutionary branches.

Most recent clusters¹⁰ are represented with flowchart symbols and points on the tree where null mutations (at marker 425) occurred are depicted with colour filled labels. A simple speculative linear timeline has been added starting with an average date of birth for the testers (1950) and reaching back to when it was thought the L21+ mutation occurred.

It can be seen that the 35 reference Collas form a neat core in the schematic and were the last to branch from the central stem in about A.D. 200. All the remaining haplotypes branched earlier which was to be expected as they had higher GDs to the Colla modal and to each other. In other words, the observed nesting was consistent with expectation.

7.2.1 Null Events

In all, the results from this dataset suggest four or five null events occurred at marker 425. In descending order they were:

1. A 9/0/22 event which was the root of the Colla population.
2. Descended from this was a 10/0/22 (219 McDougall) which was probably a simple mutation of 511 going from 9 to 10. This is supported by its place in the Phylogenetic tree and that 219 McDougall has DYS 505=9. The probability is high that this grouping is Colla.
3. A 10/0/23 event giving rise to Montero.
4. Possibly a second 9/0/22 event (classed as A) giving rise to 393McClain. This could have occurred in a different lineage about 350 years before the identical Colla one (which would be indistinguishable from it). However, as 393McClain is R-DF21 and DYS 505=9, he has to be regarded as Colla and is simply out of place in the phylogenetic tree (because of a high GD).
5. A 10/0/23 event giving rise to McQuilken.
6. A 10/0/23 event giving rise to Johnson.

None of the five testers carrying these nulls showed any degree of relatedness to each other with the exception of 219 McDougall who was part of the Colla group. The 219 McDougall/648White cluster was derived from a single lineage and can be explained by the mutation 9/0/22-23 going to 10/0/22-23.

7.2.2 Precursors and Distant Kin

The phylogenetic tree suggests the immediate precursors to the Colla Null gave rise to Carroll (L) and Mahan (H) clusters¹¹ and two generations further back, to another Carroll cluster¹². Further back still, the precursors gave rise to Williams, Bush, McKenzie. Before that, precursors for names such as Mears, Lyons, Burden, Williams, Guthrie, and Kelley appeared.

Interestingly, the root of the sample would appear to have been 9/12/23 which persisted for about 400 years before mutating to 10/12/23. From then on, 10/12 and 9/12 alternated frequently. The earlier branches certainly reflect the wide Diaspora of R-L21. The earliest were De Masci (France), Thompson (England), Basile (Sicily), Stoney (England); later, a 10/12/23 grouping consisting of 655Duffy and a cluster of N30Maguire, N20Maguire, 124McMahon, 142McMahon, 143Maguire, 312McCown whose lineage branched about 1400B.C (top right of Figure 2). This cluster is interesting in that the common ancestor (of the cluster) was about 300-400 years ago and it shows common ancestry within this lineage for McMahon, Maguire and McCown¹³. In every way, this was a cluster of surnames typically

¹⁰ Typically, those started by a common ancestor about 300-400 years ago.

¹¹ N85Carroll (L), 208McCall (L) and N19Mahan, 258McMahan, 134Corbit.

¹² 107Carroll (L), 112Carroll (K), 140Carroll (J).

¹³ This might account for the claim made by Livingstone that McMahon and McGuire are related.

associated with Clan Colla except for the absence of the null at 425 and very large GDs to the Colla modal. Some members of this cluster had the L513 and/or L69 SNP (Table 5) while others showed close relationship to the Airghialla 2 Group.

Other less distant branches of British and European origin were Montero (Spain) with null #2, McClain (Scotland) with possibly null#3, McQuilken (Scotland) with null #4; Johnson (England) with null #5, Bush (England), Borchert (Germany), Bodin (France), Hamon (France), Horwill (England), Thomas (Wales), Davis (Wales), Price (Wales), Burden (England), Williams (Wales), Rogers (England), Seville (Unknown). All the countries where R-L21+ is thought to have migrated through are represented by today's testers.

Branching (in descending order) which gave rise to Scottish names were McCall, McKenzie, McClain, McQuilken and Duncan. Those giving rise to Welsh names were Williams, Price, Morgan and Davis.

7.3 PHYLIP from Second Experiment

The data set used in the second experiment is shown in Annex A1. All the marker data was used in generating the 'Infile' for the Kitsch programme but only values for markers DYS 511, 425 and 413 are shown in Annex A1. 73 Colla and 50 non-Colla taxa were used to generate the PHYLIP output from a run with 99 iterations (jumbles). This output is shown in Annex B2.

7.3.1 Phylogram

The coloured symbols shown in Table A1 help define the boundaries and are shown in the Phylogram (Figure B2). The amber triangles represent Collas (9/0 for 511/425), yellow triangles those that are 10/0 (for 511/425) and not generally part of the Colla population with the exception of McDougall/White (see 7.2.1). Non-nulls are represented by circles, their colour coding indicating values for markers 511, 425, 413 (Table A1).

Branching does not indicate clustering as these profiles represent lineages (which in many cases in turn represent clusters) but simply where they diverged from each other. Most lineages had diverged by about the time of surname adoption.

A period of rapid branching of the Collas between c. 1750-1550 yBP¹⁴ is consistent with an invasive force taking over territory and populating it (as written records claim); this is followed by a consolidation period of less frequent branching where individual lineages established themselves in whatever territory they were in followed by the adoption of surnames.

7.3.2 Naming

The adoption of surnames doesn't map well onto the DNA evolutionary tree. This is evident among the Collas and also the non-Colla R-L21. For example, the lineages N16Calkin (English) and 368Paden (English) share a common ancestor who was also the ancestor for N49Heaney to 1820Guin (Irish and Scottish). A wider example is the common ancestor for 855Roderick and the group bounded by 145 McMahan-394McMahon. An explanation for these observed disparities could be that a close relative of Calkin or Paden migrated to Ireland &/or Scotland where they evolved into separate lineages and acquired the appropriate surname when the time came. Similarly, the ancestor who gave rise to Roderick/Rice migrated to Wales while a close relative migrated to Ireland/Scotland accounting for 63% of all Collas (inclusive of cluster members). Not so easy to explain are names such as Judd and Goodwin who are very definitely Colla and reputed to be of Anglo-Saxon origin, unless a name change by adoption or other means took place, probably at a much later date.

Among the non-Collas (represented by circles), typical Oriel Celtic names such as Carroll (including Ely Carrolls), McMahon, McGuire, Caine were found. They were descended from Colla precursors (i.e. Colla kinsmen) and tend to be 10/12 (blue circles). There is a strong representation of Welsh names among the 9/12 (mauve circles).

¹⁴ yBP= years before present. The present is taken as 1950 which approximates to dates of birth of testers.

DF 21	GD to	DURRQ	24	14	11	11-15	13	13	29	17	9-10	25	28	15-15-17-17	11	19-23	16	15	18	19	36-37	12	12	15-16	10	9	0**	22-23	16	12	16	8	12	20	13	12	9
648Higgins	9	25	14	10	11-15	14	13	29	17	9-10	25	27	15-15-17-17	11	19-23	16	15	18	17	37-37	12	12	15-16	10	9	0	22-23	16	12	16	8	12	20	13	12	DYS492	
156L21proj	11	24	14	11	11-15	14	13	29	17	9-10	25	27	15-15-17-17	11	19-23	16	15	18	17	37-37	12	12	15-16	10	9	0	22-23	16	12	16	8	12	20	13	12	DYS446	
145687 McMahon*	4	24	14	10	11-15	13	13	29	16	9-10	25	28	15-15-15-17	11	19-23	16	15	18	19	36-37	12	12	15-16	10	9	0	22-23	16	12	16	8	12	20	13	12	DYS20	
N88161 Connolly*	6	24	14	11	11-15	14	13	29	16	9-10	25	29	15-15-17-17	11	19-23	16	15	18	19	36-37	12	12	15-16	10	9	0	22-23	16	12	16	8	12	20	13	12	DYS481	
23171 McGuire*	7	24	15	11	11-15	13	13	29	17	9-10	25	28	15-15-17-17	11	19-23	16	15	18	19	37-39	12	12	15-16	10	9	0	22-23	16	12	16	8	12	20	13	12	DYS444	
N52848 Rogers*	9	24	14	10	11-15	14	13	29	16	9-10	25	27	15-15-17-17	11	19-23	16	15	18	19	36-38	12	12	15-16	10	9	0	21-23	16	12	16	8	12	20	13	12	DYS450	
21971 MacDougall*	8	23	14	10	11-15	14	13	29	17	9-10	25	28	15-15-17-17	11	19-23	16	15	18	19	36-38	12	12	15-16	10	10	0	22-23	16	12	16	8	12	20	14	12	DYS457	
125570 Connley	13	24	14	11	11-16	12	13	30	17	9-9	25	29	15-15-17-17	11	19-23	16	15	18	17	35-38	12	12	15-15	10	9	0	22-23	16	12	16	8	12	20	14	12	DYS413	
27818 Shannon	8	25	14	10	11-15	13	13	28	17	9-10	25	27	15-15-15-15	11	19-23	16	15	18	17	36-36	12	12	15-16	10	9	0	22-23	16	12	16	8	12	20	13	12	DYS425	
39313 McClain*	11	24	14	11	11-15	13	13	29	17	9-9	25	27	15-15-17-17	11	19-23	16	15	18	17	37-37	12	12	15-16	10	9	0	22-23	16	12	16	8	12	20	13	12	DYS511	
179484 McCain	14	24	14	11	11-15	14	13	29	17	9-10	25	29	15-15-17-17	11	19-23	16	15	18	19	37-37	12	12	15-16	10	9	0	22-23	16	12	16	8	12	20	13	12	DYF406S1	
209212 Coleman	13	23	14	11	11-15	13	13	29	16	9-10	25	28	15-15-17-17	11	19-23	16	15	18	17	35-38	12	12	15-16	10	10	12	23-23	16	12	16	8	12	20	14	11	DYF395S1	
208276 McCall	12	24	14	11	11-14	11	13	29	16	9-10	25	29	15-15-17-17	11	19-23	16	15	18	17	35-38	12	12	15-16	10	10	12	23-23	16	12	16	8	12	20	13	12	DYS438	
145178 Griffe*	14	24	14	11	11-14	12	13	29	15	9-10	25	28	15-15-17-17	11	19-23	16	15	18	19	37-38	12	12	15-16	10	10	12	23-23	16	12	16	8	12	20	13	12	DYS442	
122847 OMoore*	14	24	15	11	11-14	11	13	29	17	9-10	25	30	15-15-16-17	11	19-23	16	16	18	18	34-39	12	12	15-16	10	10	12	23-23	16	12	16	8	12	20	14	12	DYS456	
75621 Harris*	15	24	14	11	11-14	12	13	29	17	9-10	25	29	15-15-15-16	11	19-23	16	14	17	17	37-37	12	12	15-17	10	10	12	23-23	16	12	16	8	12	20	13	12	DYS460	
185954 Carroll*	12	25	14	11	11-14	12	12	28	16	9-10	25	26	15-15-17-17	11	19-23	16	15	18	18	37-38	12	12	15-16	10	10	12	23-23	16	12	16	8	12	20	13	12	DYS464	
76171 Cain*	15	24	14	10	11-14	12	13	29	17	9-10	25	28	15-15-17-17	11	19-23	16	15	18	19	37-37	12	12	15-16	10	10	12	23-23	16	12	16	8	12	20	13	12	DYS490	

Figure 3. Marker values for the most distant Collas and their closest non-Colla relatives. Infills are where markers differ from DURRQ. Amber text is Colla and blue text is non-Colla.

7.3.3 DNA Profiles

A qualitative examination of the Colla and non-Colla DNA profiles in general showed large differences in DYS markers, more than might be expected for a transition from non-Colla DNA to Colla DNA. However a closer examination of those profiles at their interfaces and at middle distances, showed a more gradual transition. It can be assumed that at the time of transition, all that was necessary to go from being a pre-cursor to Colla was the null mutation at 425 in 9/12 stock (Figure 2). Nevertheless, a large number of additional mutational steps would have occurred to bring about the observed changes which can be accounted for by these lineages having evolved separately since the split. The best fit to a Colla pre-cursor would be either 209 Coleman or 179 McCain. Figure 3 shows the profiles of those termed interface and middle distant (in terms of the Phylogenetic tree); only markers that showed variability are depicted.

Despite earlier reservations, neither 125 Conley (GD=13) nor 393 McClain (possibly misplaced in tree) seem out of place. Significant observations were the values for DYS 505 where all the Collas who have tested so far have a value of 9; the non-Collas are mainly 12 but range from 11-13. The lack of variability for this marker among the Collas would suggest a stable marker which must have mutated to 9 shortly before the Colla line began to evolve (by acquiring the 425 null) as there are no intermediate values (so far) among the Collas.

7.4 DF21 and Other SNPs

DF21 is an SNP downstream of the L21 SNP and its presence or absence is still being tested for. Its occurrence pre-dates the defining DYS 425 null mutation for Colla¹⁵ and could be useful in further defining Colla. So far, all Collas who have tested have proved to be DF21+ indicated by * in Figure B2; DF21- is indicated by # in Figure B2.

The Earliest DF21 found in the present dataset are the non-Collas, 761Caine (c. 2,700 yBP) and 756 Harris (c. 2,550 yBP). Ten non-Colla R-L21 have tested negative for DF21 but no inference can be drawn from their relative positions in the Phylogram as they are outside the R-DF21 population; only two, 437Connel and 755Morgan, are shown in Figure B2 as references. R-DF21 is classed as a sub-clade of R-L21. Collas, as currently defined, are a further sub-clade of R-DF21.

Apart from DF21, a number of additional SNPs are emerging which could be instrumental in defining further sub-clades within R-L21. The associations detected so far are summarised in Table 5.

DF21	Z253	L513 (L69)	M222
Denoted by * in Table A1.	196McQuilkan, 929Johnson	N30Maguire, N20Maguire, <i>(124McMahon),</i> <i>(142McMahon),</i> 143Maguire, <i>(312McCown)</i>	623Carroll, 600Carroll
Table A1	10/0/23-23	10/12/23-23	11/12/21-23 10/12/21-23

Table 5. Association of Specific Profiles with SNPs and those predicted (Italics). The Values for Markers 511, 425, 413 are given in the second row (as defined in footnote 6).

The L513 grouping (including some who have the downstream L69 only) are descended from a single lineage which branched early and form a distinct cluster. Several of them show close affinity to the Airghialla 2 modal (Ref 6).

¹⁵ MRCA estimates by Mike Walsh using Ken Nordtvedt's Generations5 spreadsheet suggest that DF21 is comparatively quite old, with a tMRCA of 3112 yBP versus 4135 yBP for all of R-L21.

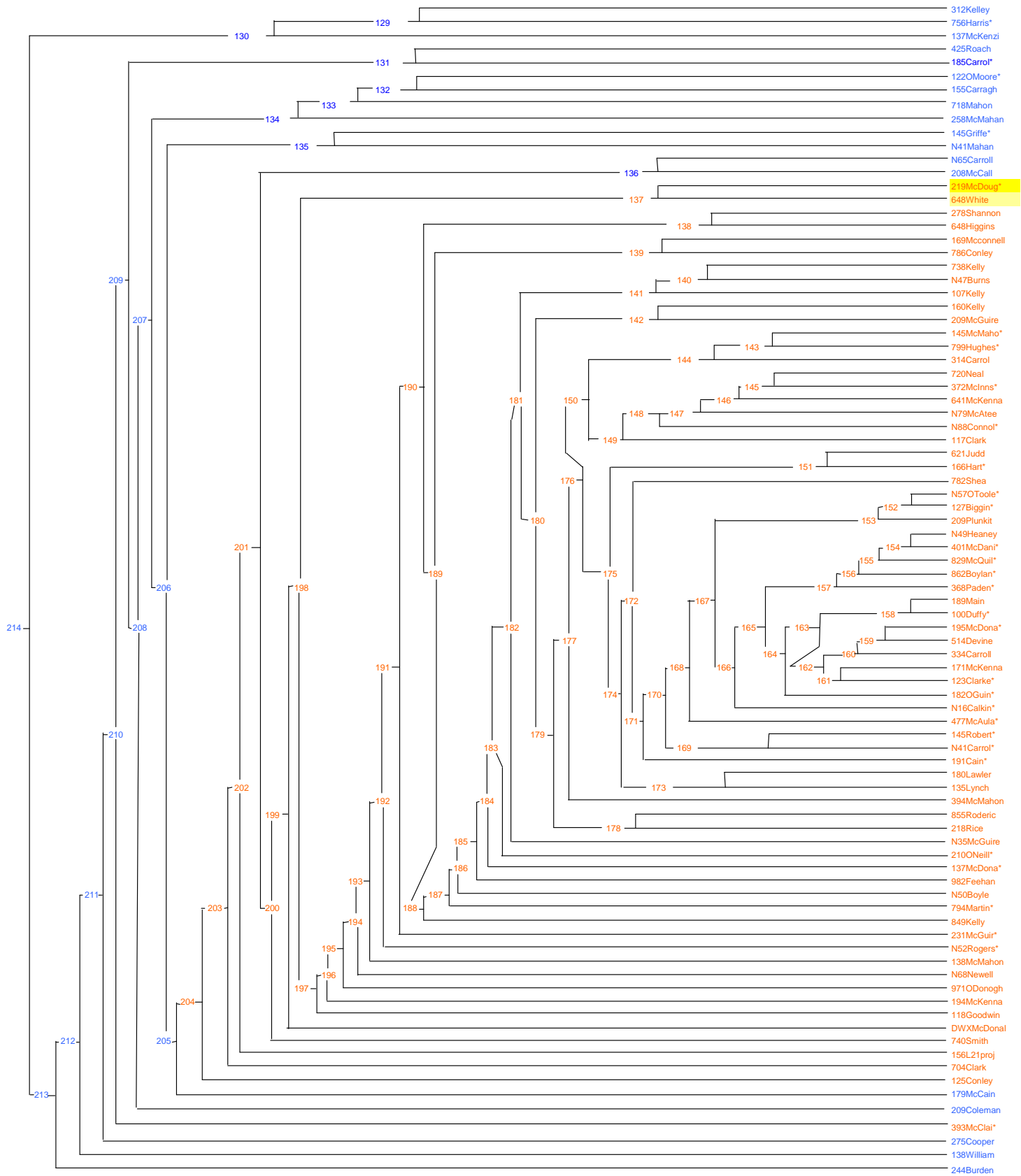


Figure 4. Schematic of the PHYLIP output for Colla and non-Colla taken from Annex C2.

7.5 Reduced Tree

The PHYLogeny Inference Package generates its output in tabular form as shown in Annex C2 from which the evolutionary tree (Phylogram) in Figure B2 was constructed (by the software). Using the data in Annex C2, a subset of the tree was constructed manually concentrating on Colla and their nearest non-Colla neighbours (Figure 4). This reduced tree is colour coded for DYS 425 nulls (amber), non-nulls (blue) and a yellow background for the 10/0 configuration of the McDougall/White lineage; also shown are the software deduced ancestors given in Annex C2.

Overall, the level of symmetry is good with the Collas bounded by 125Conley on one side of the tree and 219McDougall on the other. The exception is the interface between the N65Carroll /208McCall lineage and the Collas. In an earlier experiment (not shown), this lineage was well separated from the Collas but 125Conley came out as a separate null event. In this data set, 125 Conley appears to be the earliest branch from the Colla origin despite its GD of 13 from the modal. Applying the average rate of 167 years/mutation (sec 6) to 125 Connolly with a GD of 13, gives a time estimate of 2171 yBP for the earliest detectable Colla. 393McClain shows up as a separate 425 null event re-enforcing the same finding from experiment 1 (7.2.1).

7.5.1 GD Net Sum

The dilemma posed by GD and its cut-off point is apparent from this dataset. STR mutations, like all mutations, will occur in both forward and reverse directions at different rates. It is possible in a limited number of cases to track forward and reverse mutations (using a tool such as Fluxus) but for the bulk of data processed through the McGee Tool (as in this analysis), the deduced GD values will be an aggregate of forward and reverse mutations, a GD net sum. If for example, the total number of mutational differences (from a reference) is 13 and no reversion has occurred, then it's observed GD will be 13; if one of the forward mutations reverses, then the observed GD will be 11. Likewise, with 11 mutational events and one reversion, the observed GD will be 9. This is the most likely explanation for 125Connolly (a very obvious Colla) having a GD of 13 (from DURRQ) and appearing very early in the phylogram (which is driven by GD in the form of tMRCA values). It would also be the possible explanation for N65Carroll /208McCall appearing later than expected for non-Collas.

7.5.2 Common Ancestry

An expansion of Figure 4 (to its right) is shown in Figure 5. This is where the clusters associated with lineages are added in to the assessment. The bulk of the profiles are towards the core of the evolutionary tree which might suggest that these are the most successful Colla lines (in terms of propagation) but it also reflects the non random DNA testing¹⁶. Ancestor 150 is responsible for 27 profiles (145McMahon to 117Clark) and ancestor 172 for 74 profiles (782Shea to 191Cain). These are effectively combined (plus extras) under ancestor 177, accounting for 110 profiles (62.5% of all Collas). Expanding further will account for more Collas. The earliest ancestor who accounts for 98.3% of Collas is ancestor 200 which excludes 125 Conley, 704Clark and an unknown. Earlier ancestors are common to Collas and non-Collas alike, the separation appearing to be at ancestor 205 giving rise to non-Colla 179McCain and the bulk of Collas.

7.6 Surnames and Geography

This is the most difficult aspect of the analysis. The comfortable aspect of genealogy that most would like to see is that all the bearers of a particular surname would share a common ancestor in DNA terms. Unfortunately, this is not the case in the current study (or in other projects as judged by their name diversity). Surname diversity can be broadly categorised under the following headings.

7.6.1 Pre-surname

Most of the evolution examined in the current analysis took place long before the introduction of surnames. The schematic in Figure 6 is an attempt to draw together the salient features of the present

¹⁶ Certain family groups persuaded known and suspected relatives to test leading to some large clusters. Lineages by and large should be free of this influence.

study. This pictogram attempts to show the Collas in context with the larger tribal groups and the approximate timing of the different strategic events. There is good archaeological evidence for the movement of Celtic people through Europe, Britain and Ireland for the timeframe depicted. This population or tribal drift would have been a continuous process over time with the occasional invasive event. It is reasonable to assume that the R-L21 clade (and later the R-DF21 sub-clade) was a major component in this tribal drift since 60-70% of all R-L21 are to be found in Ireland today¹⁷. Technically, all those carrying the L21 SNP are related, albeit most very distantly given the time span. This relationship becomes progressively closer in the sub-clades, culminating in the sub-sub-clade defined as Colla.

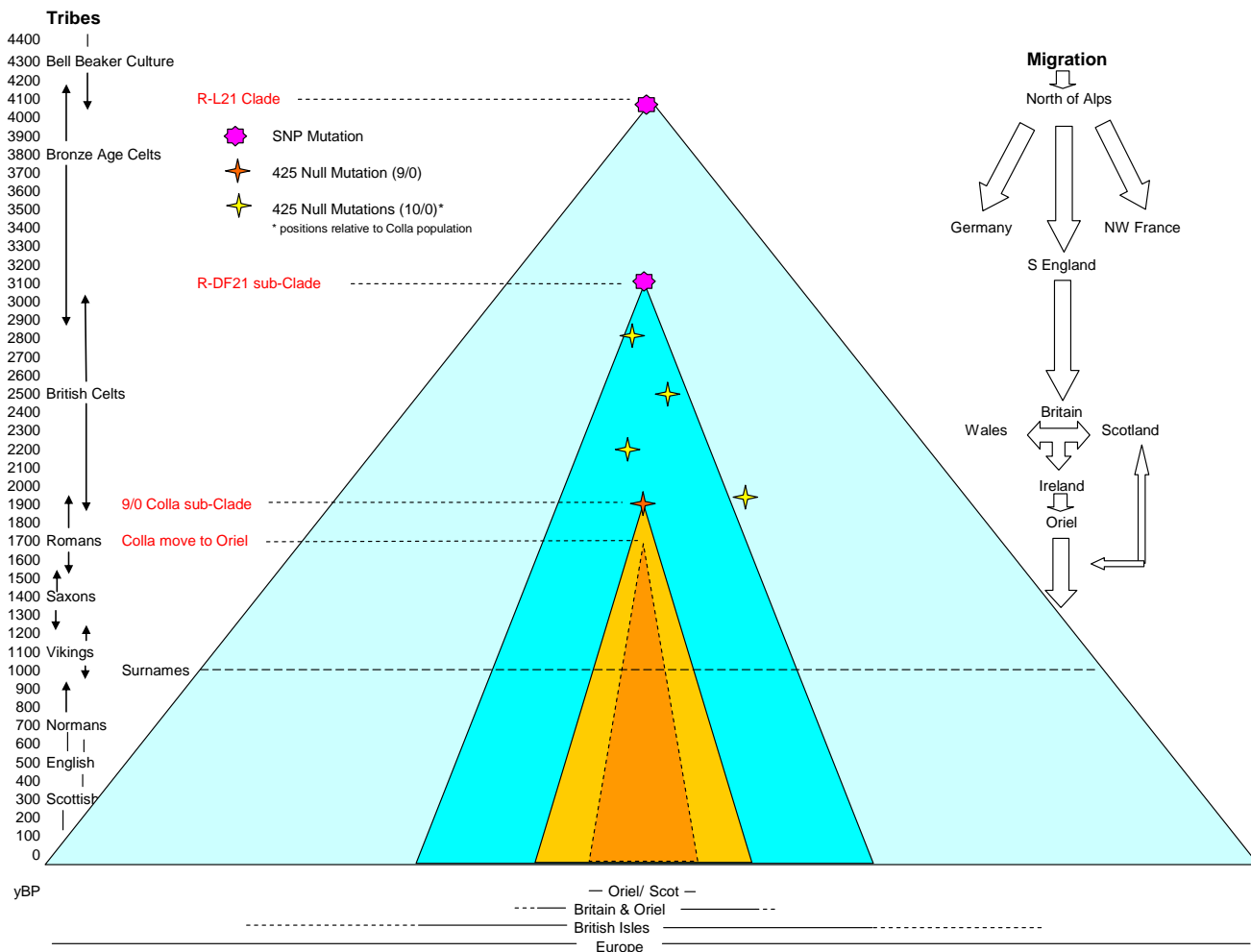


Figure 6. Pictogram relating key events in the evolution of the Celtic R-L21 Clade.

7.6.2 Colla Origins

There will always be controversy over whether the roots of Collas with a strong Scottish, Welsh or English presence (today) were indigenous to mainland Britain or migrated from Ulster to mainland Britain¹⁸. The latter would be counter to the population drift of the Celtic tribes in general, and it would have been difficult to establish so many lineages over such a wide area. The Results would suggest that Collas were well established in Britain before conquering Oriel. The converse (excluding much later 18/19th c. migrants), although not impossible, would have been more difficult to achieve as it would imply a large invasive force conquering and holding territory so as to have been able to propagate over several generations. The exception would have been Alba/DalRiada (which was a unified Kingdom on both sides of the Irish Sea) and from which a large proportion of Collas seem to have originated; indeed, the evidence is suggestive that the Collas might well have originated in Alba.

¹⁷ According to the Co-admin of the "Ireland yDNA" Project.

¹⁸ The spread of Collas seems to have been more widespread on Mainland Britain than in Ulster, where they are largely limited to the Oriel territory.

7.6.3 Geographic.

Even though the DNA was of Celtic origin, the place of residence (and the chieftain to whom you owed loyalty) at the time of naming would have dictated the type of surname adopted. For example, if living in a Saxon controlled (but formerly Celtic) area such as Wessex, it would not be surprising if Saxon names were adopted; likewise, in Norman areas later, adoption of Norman names would have been natural. The same argument can be applied to Scotland, Wales and Ireland. Most of the Colla lineages (73 in the present study) had devolved prior to surnames being adopted and it is expected that these would have travelled throughout mainland Britain, if not the British Isles, during this lengthy time period.

7.6.4 Adoption

Many are sensitive to having an adopted name while others are happy to admit to it as it gives them some sense of identity knowing their DNA is of Colla origin (for the present examples). There are a few known adopted names (sec 6.3) and many more which are suspected. Prior to the introduction of surnames, patronymics of one sort or another were used by Celtic tribes which only had local significance. There were many other reasons for name changing other than by adoption. In the years following the Norman invasion of Ireland there was a fair amount of fraternisation which on occasion could have resulted in progeny with Colla DNA having Norman names¹⁹; this would have continued to the present day. Later, there was active suppression (Penal Laws) of Gaelic names leading to names after places, colours, objects etc. Lack of registration until the mid 19th c. meant that name changing for various social reasons would have been easy. Proof of adoption has only been possible since the introduction of institutions and associated documentation in the 19th c.

8 Discussion

These results show that the Collas were not a stand-alone clan but were part of a much larger Celtic grouping, a sub-set in fact which in all probability can be defined by 9/0/9 values for markers DYS 511, 425, and 505. Where and approximately when they arose is clearly shown but not why or how they proliferated, as clearly they must have done, compared to other instances of null mutations at marker 425. A rational explanation for their expansion is that they (as a family group) must have enjoyed privileges²⁰ which allowed them to procreate prolifically over the first few generations (after the null mutation) in order for their descendants to be so numerous today (approaching 20% of the R-L21 population). Most lineages had diverged by about the time of surname adoption. A period of rapid branching of the Collas between c. 1750-1550 yBP is consistent with an invasive force taking over territory and populating it (as written records claim); this is followed by a consolidation period of less frequent branching where individual lineages established themselves in whatever territory they were in followed by the adoption of surnames.

Admittedly, the large number of Collas among today's testers is heavily influenced by large groups of family members (whose common ancestors were c. 200-300 years ago) and in terms of the longer timescale being examined here, would be classed as single lineages. Equally, many of those who are single lineages down to today are probably so because either known family members have failed to test or they are the only members to have survived into the 20th century. Assuming testing is fairly random, the actual number of lineages presenting today will generally reflect the size of the population from which they are descended. On this basis, the starting Colla population must have grown enormously shortly after it arose to be so well represented in the 20th century.

The close branching observed in the evolutionary tree roughly between 2000-3000 yBP could represent the rapid expansion of the R-L21 Celts through NW Europe and Britain. Later, a similar expansion was observed for Collas which would support the historical view that it was a specific event rather than part of a general migration.

¹⁹ This would also have applied in Britain, post A.D.1066.

²⁰ As there is no known biological advantage to having a null mutation at 425, the 'privilege' must have been of a social nature such as high numbers of offspring enjoying high survivability over several generations. This would be expected for a family of a dominant chieftain who had conquered a territory, as history describes for the Colla brothers.

Although the testers examined are widely separated in terms of STRs, in conventional genetics, they would be regarded as closely related. They all share the SNP L21+ and some share the SNP DF21+, whose mutation rates are measured in thousands of years, whereas the 67 STRs examined have an approximate rate of one mutation per 167 years. DF21 is an SNP downstream of the L21 SNP. All Collas who have tested for this marker have proved to be DF21+. The Earliest DF21+ found in the present dataset are the non-Collas, 761Caine (c. 2,700 yBP) and Harris (c. 2,550 yBP). R-DF21 is classed as a sub-clade of R-L21. Collas, as currently defined, would appear to be a further sub-clade of R-DF21.

This closeness in relationship is demonstrated by having Colla names such as Carroll, McMahon, McGuire, Duffy, Kelly, as non-Colla and Colla lineages. This view is especially endorsed by the very early (2800 yBP) lineage which gave rise to a classical Colla name grouping of McMahon, Maguire and McCown (a cluster of six testers in all), 1600 years later. Members of this grouping have the SNP L153 and/or L69 as well as showing close affinity to the Airghialla 2 modal.

This leads into surnames and how they were acquired by null and non-null alike around A.D. 1000. The adoption of surnames hasn't mapped well onto the DNA evolutionary tree. This is evident among several lineages that share common ancestors and give rise to a wide variety of names. Among the non-Collas, typical Oriel Celtic names such as Carroll (including Ely Carrolls), McMahon, McGuire, and Caine are found so the genetic origin of these names is by no means exclusive to the Colla line. An explanation for these observed disparities could be that close relatives migrated to Ireland &/or Scotland &/or Wales where they evolved into separate lineages and acquired the appropriate surname when the time came. A simplistic view might be that a Colla Chieftain would have bestowed on those he thought were his family and relatives (and possibly some family retainers) the name he had become known by. This would have had to happen in all of the twenty or so Colla clans which would seem improbable given the nature of selecting clan chiefs under Brehon law²¹. Given the DNA picture, a much more likely process would have been that a mixture of null and non-null tribal members accompanied the Colla brothers to Oriel where they obliterated the indigenous people and established themselves in the territory where they would have proliferated over the next few hundred years. Important families of null and non-null alike would have enjoyed 'privileges' (as referred to above) but perhaps the most privileged (judging by today's numbers) would have been null descendants of the original Colla leaders. By the time naming was adopted, descendants of the Colla invasion force, both null and non-null, would have been randomly distributed throughout their conquered territory of Oriel. Consequently when naming took place, the recipients of the McMahon, Carroll, and McGuire etc names would have been a mixture of null and non-null people.

In addition to the Irish lineages, a similar pattern of null and non-null Scottish and Welsh lineages were also observed which could account for the null and non-null members of these clans. Previous studies showed that about half of Colla lineages were of Scottish origin and two were of Welsh origin. The present results show at least five non-null Scottish lineages and three representing the Collas; there were at least four non-null Welsh lineages and two representing the Collas. Although possible, it is hard to envisage the early Colla sub-set being so widely dispersed, from S Wales to the Scottish Isles, but could have occurred if the latest estimate of the null mutation occurring about 2000 yBP is approximately correct. On the other hand, the geographic location of ancestors 3 or 4 generations back from today may have no significance; it may simply have been the destination of much later migrations, including those from Ireland in the 18/19th century. The high representation of Scottish clans among the Collas is suggestive that the mutation may have occurred in Scotland although this could equally well be accounted for by the historically recorded movements between Ulster and Dal Riada. From the observed results, the balance of probability is that the Colla population established itself in mainland Britain before some lineages migrated to Oriel. This could account for some of the earlier branches in the evolutionary tree giving rise to some English (Saxon, Norman) type names with later ones (of McDaniel, McDonald) evolving from the returned Colla Uais and his retinue to Alba.

²¹ Under Brehon Law, there was no automatic line of succession to sons but rather there would be lots of claimants (out to 2nd cousins) and probably many illegitimate ones as well. Essentially, claimants fought one another for the title and the successor was the successful combatant.

9 Conclusions

The conclusions that can be drawn from this study were:

- The Colla population started with a specific null mutation in marker DYS 425 in a lineage that had a value of 9 for marker DYS 511. The suggested timing of this event from this analysis is about the beginning of the first millennium which is consistent with previous estimates and the known history for the period.
- All the remaining non Colla haplotypes in the dataset branched earlier than Colla and their relatedness was proportional to the time of branching.
- The defined Colla population is a distinct subset of R-DF21 which in turn is a sub-set of R-L21.
- Other nulls occurred at marker DYS 425 independently of each other and at different times usually with a value of 10 at marker DYS 511; 3/4 of these were non-Colla.
- Colla can be defined as R-L21, R-DF21, DYS 511=9, DYS 425=0, and potentially DYS 505=9.
- Non-Colla DNA lineages evolved in parallel with Colla DNA and acquired a similar range of Colla names.
- The spread in the Phylogenetic tree is consistent with the geographic spread of R-L21 throughout Europe, Britain and Ireland.
- Colla DNA and non-Colla DNA alike spread throughout the British Isles and acquired clan names, appropriate to their geographic location and/or local chieftain, at the beginning of the second millennium.

10 References

1. Analysis of Colla DNA Profiles (Draft 2), June 2010Patrick McMahon
2. McMahon DNA (Draft), April 2011Patrick McMahon
3. http://www.eupedia.com/europe/maps_Y-DNA_haplogroups.shtml#R1b-L21
4. McGee's Calculator..... <http://www.mymcgee.com/tools/yutility111.html>
5. [http://home.earthlink.net/~odoniv/HamCountry/HAM_DNA_Project/HAM_DNA_Phylogenetic chart Instructions.html](http://home.earthlink.net/~odoniv/HamCountry/HAM_DNA_Project/HAM_DNA_Phylogenetic_chart_Instructions.html)
6. Personnel communicationPeter Biggins

A. Input Data

A.1. Data used in Experiment 2.

GD	Lineage Rep	Rest of Cluster	ID	Label	DF 21 * = Pos; # = Neg	DYS511	DYS425	DYS413	Symbol	
					DURRQ	9	0**	22-23		
6	127Biggins	N86Bigham	127469	A	127Biggin*	9	0	22-24	▲	
5			N86783		N86Bigham	9	0	22-24		
4			146Biggins	146867		146Biggins	9	0	22-24	
4			N34Beaghen	N34030		N34Beaghen	9	0	22-24	
6			696Little	69648		696Little	9	0	22-24	
7			166Beggan	166169		166Beggan	9	0	22-24	
5	N16Calkins	569Calkins	N16767	B	N16Calkin*	9	0	23-23	▲	
5			56933		569Calkins	9	0	22-23		
4			173Calkins	173760		173Calkins	9	0	22-23	
4			176Calkins	176713		176Calkins	9	0	22-23	
6			150Calkins	150549		150Calkins	9	0	22-23	
7			104Calkins	104240		104Calkins	9	0	22-23	
6	N41Carroll	N41845	N41845	C	N41Carrol*	9	0	22-23	▲	
4			167Carroll	167430		167Carroll	9	0	22-23	
6			176Carroll	176233		176Carroll	9	0	22-23	
5			306Carrol	30624		306Carrol	9	0	22-23	
6			320Carroll	32061		320Carroll*	9	0	22-23	
7			610Hendric	61096		610Hendri*	9	0	22-23	
3	334Carroll		E10Dundas	33486	D	334Carroll	9	0	22-23	▲
4		E10544	E10544		E10Dundas*	9	0	22-23		
6	314Carrol		31491		314Carrol	9	0	22-23	▲	
8	N50Boyle		N50500		N50Boyle	9	0	22-23	▲	
4	862Boylan	652Boylan	86258	E	862Boylan*	9	0	22-23	▲	
3			65227		652Boylan	9	0	22-23		
4	123Clarke	117758Clar	123611	F	123Clarke*	9	0	22-23	▲	
4			130Clarke	130308		130Clarke	9	0	22-23	
7			117758Clar	117758		117758Clar	9	0	22-23	
5	117Clark		117759	G	117Clark	9	0	22-23	▲	
1		628Hancock	62858		628Hancock	9	0	22-23		
2	514Devine	117Devine	51491	H	514Devine	9	0	22-23	▲	
6			117316		117Devine	9	0	22-23		
5		166McDanie	16694		166McDanie	9	0	22-22		
3		684McDanie	68473		684McDanie	9	0	22-23		
3	401McDanie	N18Alexand	40141	I	401McDani*	9	0	22-23	▲	
3			169McDanie	169145		169McDanie	9	0	22-23	
5		N18Alexand	N18664		N18Alexand	9	0	22-23		
6	N88Connoll		N88161		N88Conno*	9	0	22-23	▲	
6	786Conley		78625		786Conley	9	0	22-23	▲	
6	169Mcconnell		169972		169Mcconnell	9	0	20-23	▲	
6	135Lynch		135626		135Lynch	9	0	22-23	▲	
5	191Cain		191286		191Cain*	9	0	22-23	▲	
8	982Feehan		98203		982Feehan	9	0	22-23	▲	
9	971ODonogh		97145		971ODonogh	9	0	22-23	▲	
4		374Kern	37416		374Kern	9	0	22-23		
4	N49Heaney		N49738	J	N49Heaney	9	0	22-23	▲	
7	209McGuire		209574		209McGuire	9	0	22-23	▲	
8	218Rice		21802		218Rice	9	0	22-23	▲	
6	N79McAtee		N7992		N79McAtee	9	0	22-23	▲	
6	180Lawler		180119		180Lawler	9	0	22-23	▲	

5	209Plunkit		209101		209Plunkit	9	0	23-23	▲
6	N47Burns		N47257		N47Burns	9	0	22-23	▲
6	372McInns		37243		372McInns*	9	0	22-23	▲
5	182OGuin		182242		182OGuin*	9	0	22-23	▲
8	782Shea		78217		782Shea	9	0	22-23	▲
9	N52Rogers		N52848		N52Rogers*	9	0	21-23	▲
6		984Duffy	98477		984Duffy	9	0	22-23	
3		N81Duffy	N81681		N81Duffy	9	0	22-23	
5	100Duffy		100022	K	100Duffy*	9	0	22-23	▲
8		128Godwin	128287		128Godwin	9	0	22-23	
10	118Goodwin		118691	L	118Goodwin	9	0	21-23	▲
9		145Godwin	145564		145Godwin	9	0	22-23	
5	166Hart		166797	M	166Hart*	9	0	22-23	▲
7		182Harte	182999		182Harte	9	0	23-23	
7		707Higgins	70725		707Higgins	9	12	22-23	
9	648Higgins		64895	N	648Higgins	9	0	22-23	▲
5		964Higgins	96415		964Higgins	9	0	22-23	
5		271Hughes	27184		271Hughes	9	0	22-23	
5		200Hughes	200241		200Hughes	9	0	22-23	
5	799Hughes		7996	O	799Hughes*	9	0	22-23	▲
4		194Hughes	194157		194Hughes	9	0	22-23	
4		119Herring	119300		119Herring	9	0	22-23	
6		178Campell	178848		178Campell	9	0	22-23	
5		116Goggins	11664		116Goggins	9	0	22-23	
5		131Almond	131057		131Almond	9	0	22-23	
4	621Judd		62180	P	621Judd	9	0	22-23	▲
5		187Judd	187811		187Judd	9	0	22-23	
7	849Kelly		84941		849Kelly	9	0	22-23	▲
8	738Kelly		73887		738Kelly	9	0	22-23	▲
6	107Kelly		107942		107Kelly	9	0	22-23	▲
8	160Kelly		160515		160Kelly	9	0	23-23	▲
8	219McDouga		21971	Q	219McDoug*	10	0	22-23	▲
8		144McDouga	144808		144McDouga	10	0	22-23	
4	477McAulay		47703	R	477McAula*	9	0	22-23	▲
5		113Beeman	113649		113Beeman	9	0	22-23	
6		112McAulay	112471		112McAulay	9	0	22-22	
4		395McCall	39559		395McCall	9	0	22-23	
2	189Main		189681		189Main	9	0	22-23	▲
11	393McClain		39313		393McClai*	9	0	22-23	▲
9		688Edwards	68859		688Edward*	9	0	22-23	
7		146McDanie	146160		146McDanie	9	0	22-23	
6		124McDonal	124947		124McDonal	9	0	22-23	
7		112McDanie	112197		112McDanie	9	0	22-23	
8		133McDonal	133546		133McDonal	9	0	22-23	
8	137McDonal		13758	S	137McDona*	9	0	22-23	▲
4		129McDonal	129206		129McDonal	9	0	22-23	
4		156McDonal	156049		156McDonal	9	0	22-23	
5		187McDonal	187107		187McDonal	9	0	22-23	
3		183McDonne	183533		183McDonne	9	0	22-23	
4	195McDonal		195306	T	195McDona*	9	0	22-23	▲
8		322McGuire	32277		322McGuire	9	0	22-23	
9		337McGuire	33735		337McGuire	9	0	22-23	
7		111McGuire	111694		111McGuire	9	0	22-23	
7	231McGuire		23171	U	231McGuir*	9	0	22-23	▲
7		N72McGuire	N7273		N72McGuire	9	0	22-23	
9		212McGuire	21218		212McGuire	9	0	22-23	
8		208McGuire	20871		208McGuire	9	0	22-23	

7		864McGuire	86421		864McGuire	9	0	22-22	
7	N35McGuire		N35961	V	N35McGuire	9	0	22-22	▲
8		108McGuire	108908		108McGuire	9	0	22-22	
7		170McGuire	170292		170McGuire	9	0	22-22	
10		158McGuire	158193		158McGuire	9	0	22-22	
9		865McGuire	86542		865McGuire	9	0	22-22	
7	794Martin		79458		794Martin*	9	0	23-23	▲
5	641McKenna		6419		641McKenna	9	0	22-23	▲
7		150Boyd	150811		150Boyd	9	0	22-23	
5		878Moore	87853		878Moore	9	0	22-23	
5		683McKinne	68390		683McKinne	9	12	22-23	
5	171McKenna		171658	W	171McKenna	9	0	22-23	▲
5		100McKenna	100576		100McKenna	9	0	22-23	
9		211McKenna	21156		211McKenna	9	0	22-23	
6		149Byrne	149290		149Byrne	9	0	22-23	
6		659McKenna	65962		659McKenna	9	0	23-23	
10	194McKenna		194128		194McKenna	9	0	23-23	▲
4	145McMahon		145687	X	145McMaho*	9	0	22-23	▲
3		N70Moorhea	N70960		N70Moorhea	9	0	22-23	
3		197McMahon	197862		197McMahon	9	0	22-23	
4		144McArdle	144005		144McArdle	9	0	22-23	
4		162Bogner	162173		162Bogner	9	0	22-23	
4		522Matthew	52246		522Matthew	9	0	22-23	
3		148McMahon	14876		148McMahon	9	0	22-23	
3		183McMahon	183423		183McMahon	9	0	22-23	
6		693McMahon	69356		693McMaho*	9	0	22-23	
6		831Smith	83132		831Smith	9	0	22-23	
8	138McMahon		13852		138McMahon	9	0	22-23	▲
7	394McMahon		39409		394McMahon	9	0	22-23	▲
4		490Woods	49098		490Woods	9	0	22-23	
4		145McQuill	145804		145McQuill	9	0	22-23	
4	829McQuill		82960	Y	829McQuill*	9	0	22-23	▲
3		170Collins	170471		170Collins	9	0	22-23	
5		209Collins	209736		209Collin*	9	0	22-23	
5		746Collins	74679		746Collin*	9	0	22-23	
4	720Neal		7206	Z	720Neal	9	0	22-23	▲
5		720Neal	7207		720Neal	9	0	22-23	
5		181Neal	181440		181Neal	9	0	22-23	
6	2100Neill		210109		2100Neill*	9	0	22-23	▲
4		112Walker	112086		112Walker*	9	0	22-23	
4		412Peden	41200		412Peden	9	0	22-23	
4	368Paden		36836	AA	368Paden*	9	0	21-23	▲
3		398Paden	39872		398Paden	9	0	22-23	
3		409Peden	40917		409Peden	9	0	22-23	
5		236Roberts	23643		236Roberts	9	0	22-23	
5	145Roberts		145674	AB	145Robert*	9	0	22-23	▲
7		4995Roderi	49959		4995Roderi	9	0	22-23	
8		8553Roderi	8553		8553Roderi	9	0	22-23	
8		49963Roder	49963		49963Roder	9	0	22-23	
11		49961Roder	49961		49961Roder	9	0	22-23	
9		49960Roder	49960		49960Roder	9	0	22-23	
9		124Roderic	12484		124Roderic	9	0	22-23	
9		185Roderic	185740		185Roderic	9	0	22-23	
8		167Roderic	167671		167Roderic	9	0	22-23	
8	855Roderic		8551	AC	855Roderic	9	0	22-23	▲
8		134Wilson	134799		134Wilson	9	0	22-23	
7		168Roderic	168082		168Roderic	9	0	22-23	
7		49962Roder	49962		49962Roder	9	0	22-23	
10		865Roderic	8651		865Roderic	9	0	22-23	

8	278Shannon		27818	AD	278Shannon	9	0	22-23	▲	
7		2782Shanno	27820		2782Shanno	9	0	22-23		
10	740Smith			74097	AE	740Smith	9	0	22-23	▲
7		836Dever	83647		836Dever	9	0	22-23		
8			119Smith	119122		119Smith	9	0	22-23	
8	N570Toole		N57121	AF	N570Toole*	9	0	22-23	▲	
10	N68Newell		N68057		N68Newell	9	0	22-23	▲	
10	DWXMcdonal		&DWXBQ		DWXMcdonal	9	0	22-22	▲	
10	648White		64804		648White	10	0	22-23	▲	
11	156L21proj		156013		156L21proj	9	0	22-23	▲	
11	704Clark		70435		704Clark	9	0	22-23	▲	
11		112Carroll	112378		112Carroll	10	12	22-23		
11			304Bowes	30445		304Bowes	10	12	22-23	
11			116Purcell	116798		116Purcell	10	12	22-23	
11			233Springer	23350		233Spring*	10	12	22-23	
12	185Carroll			185954	AG	185Carroll*	10	12	22-23	●
13		714Carroll	71400		714Carroll	10	12	22-23		
14			107Carroll		107Carroll	10	12	23-23		
12			190Carroll	190806		190Carroll*	10	12	21-23	
15			140Carroll			140Carroll	10	12	21-23	
12	208McCall			208276		208McCall	10	12	23-23	●
13	N65Carroll		N65343		N65Carroll	10	12	23-23	●	
13	209Coleman		209212		209Coleman	10	12	23-23	●	
11	N41Mahan		N41701		N41Mahan	10	12	22-23	●	
12	155Carragh		155090		155Carragh	10	12	23-23	●	
12	275Cooper		2751		275Cooper	10	12	23-23	●	
13	125Conley		125570		125Conley	9	0	22-23	▲	
13	258McMahan				258McMahan	9	12	23-23	●	
14	425Roach		42527		425Roach	10	12	23-23	●	
14	179McCain		179484		179McCain	9	12	22-23	●	
14	145Griffet		145178		145Griffe*	10	12	23-23	●	
14	1220Moore		122847		1220Moore*	10	12	23-23	●	
14	162Moore	130Moore	162107	AH	162Moore	10	12	23-23	●	
14			130051		130Moore	10	12	23-23		
14	718Mahon				718Mahon	9	12	23-23	●	
15	204Hughes		20443		204Hughes	10	0	22-23	▲	
15	756Harris	183Harris	75621	AI	756Harris*	10	12	23-23	●	
15			183983		183Harris	10	12	23-23		
15	761Caine	120Caine	76171	Aj	761Caine*	10	12	23-23	●	
14			120655		120Caine	10	12	23-23		
15	138William	147William		AK	138William	9	12	23-24	●	
18			147William			147William	9	12	23-25	
15	N59Carroll				N59Carroll	10	12	23-23	●	
16	244Burden				244Burden	9	12	23-23	●	
17	174Simons				174Simons	9	12	23-23	●	
18	110McMahon				110McMahon	10	12	23-23	●	
18	137McKenzi				137McKenzi	9	12	20-23	●	
18	N54Thomas				N54Thomas	9	12	23-23	●	
18	401Kelley				401Kelley	10	12	22-23	●	
18	312Kelley				312Kelley	10	12	23-23	●	
18	N59Duffy				N59Duffy	10	12	24-24	●	
19	196McQuilk				196McQuilk	10	0	23-23	▲	
19	130Duffy				130Duffy	11	12	22-23	●	
20	179Davis				179Davis	9	12	23-23	●	
20	173Morgan				173Morgan	9	12	23-23	●	
20	945Price				945Price	9	12	23-24	●	
21	929Johnson				929Johnson	10	0	23-23	▲	
21	E85Montero				E85Montero	10	0	23-23	▲	

21		150Mahon		150Mahon	10	12	23-23		
21	141McMahon		AL	141McMahon	10	12	23-23	●	
21	246Guthri			246Guthri	9	12	23-23	●	
21	N42Kelly			N42Kelly	10	12	21-23	●	
21	167Kelly			167Kelly	10	12	23-23	●	
22	898McMahon			898McMahon	10	12	23-23	●	
21	112Byrne			112Byrne	9	12	21-23	●	
22	600Carroll			600Carroll	10	12	21-23	●	
22	128Doherty			128Doherty	9	12	21-23	●	
22	821Duncan			821Duncan	9	12	22-23	●	
22	142McMahon		AM	142McMahon	10	12	23-23	●	
20	N20Maguire		AM	N20Maguire	10	12	23-23	●	
23		143Maguire		143Maguire	10	12	23-23		
24		N30Maguire		N30Maguire	10	12	23-23		
25		124McMahon		124McMahon	10	12	23-23		
26	312McCown		AM	312McCown	10	12	23-23	●	
25	623Carroll			623Carroll	11	12	21-23	●	
23	959Jones			959Jones	9	12	23-23	●	
24	655Duffy			655Duffy	10	12	23-23	●	
22	437Connell		43722	437Conne#	10	12	23-23	●	neg
23	755Morgan		75572	755Morgan#	10	12	23-23	●	neg
26	865Thomps			865Thomps	9	12	23-23	●	
28	173Rogers			173Rogers	9	12	23-25	●	
29	137DeMasci			137DeMasci	9	12	23-23	●	
	Amber font are nulls	Black font are cluster members					No fill are 9/0		
	Blue font are non-nulls	(not in tree)					Yellow fill are 10/0		
							Blue fill are 10/12		
							Mauve fill are 9/12		
							Green fill are 11/12		

Table A1. The selection of testers examined in Experiment 2. Testers examined in PHYLIP were those in the second column, 'Lineage Rep'. The remainders of the clusters identified in McGee are shown in the third column. The fifth column shows the identifier for each cluster. Values for the three key markers and their colour coding are shown in the 7th, 8th & 9th columns. The tenth column shows the symbols used in the Phylogenetic Tree (Annex B 2). Not shown is the remainder of the DYS data although they were used in constructing the Phylogenetic Tree.

B. PHYLIP Inputs and Outputs

The methods used in generating PHYLIP (PHYLogeny Inference Package) output are outlined in section 6.5. Only 67 marker data was used to prepare the input data. The option ‘Generate PHYLIP Data’ was checked (in addition to Hybrid model, Probability 95%, Custom mutation rate, Generations) in the McGee Y-DNA Comparison Utility. This PHYLIP data was saved as a text ‘infile’. The ‘infile’ was processed through the Kitsch software following the recommendations of Hamm (Ref 5) and the ‘outtree’ file was read using the Mega5 software. The output results are presented in the ‘Input Order’ format where the most recent branching is nested towards the centre of the layout and the most distant towards the peripheries.

B.1. Experiment 1

The 105 taxa listed in Table 3 were processed through Kitsch with 11 iterations. The time scale is arbitrary and shows the approximate years before present (yBP)²². The two further points on the time scale were the onset of the Colla null at the beginning of the first millennium and the oldest branch (so far) being a few hundred years downstream of the L21+ mutation (c. 4135 yBP). This timescale must only be used as a rough guide until more data becomes available.

²²

The ‘Present’ is taken as 1950, an approximation of dates of birth for testers.

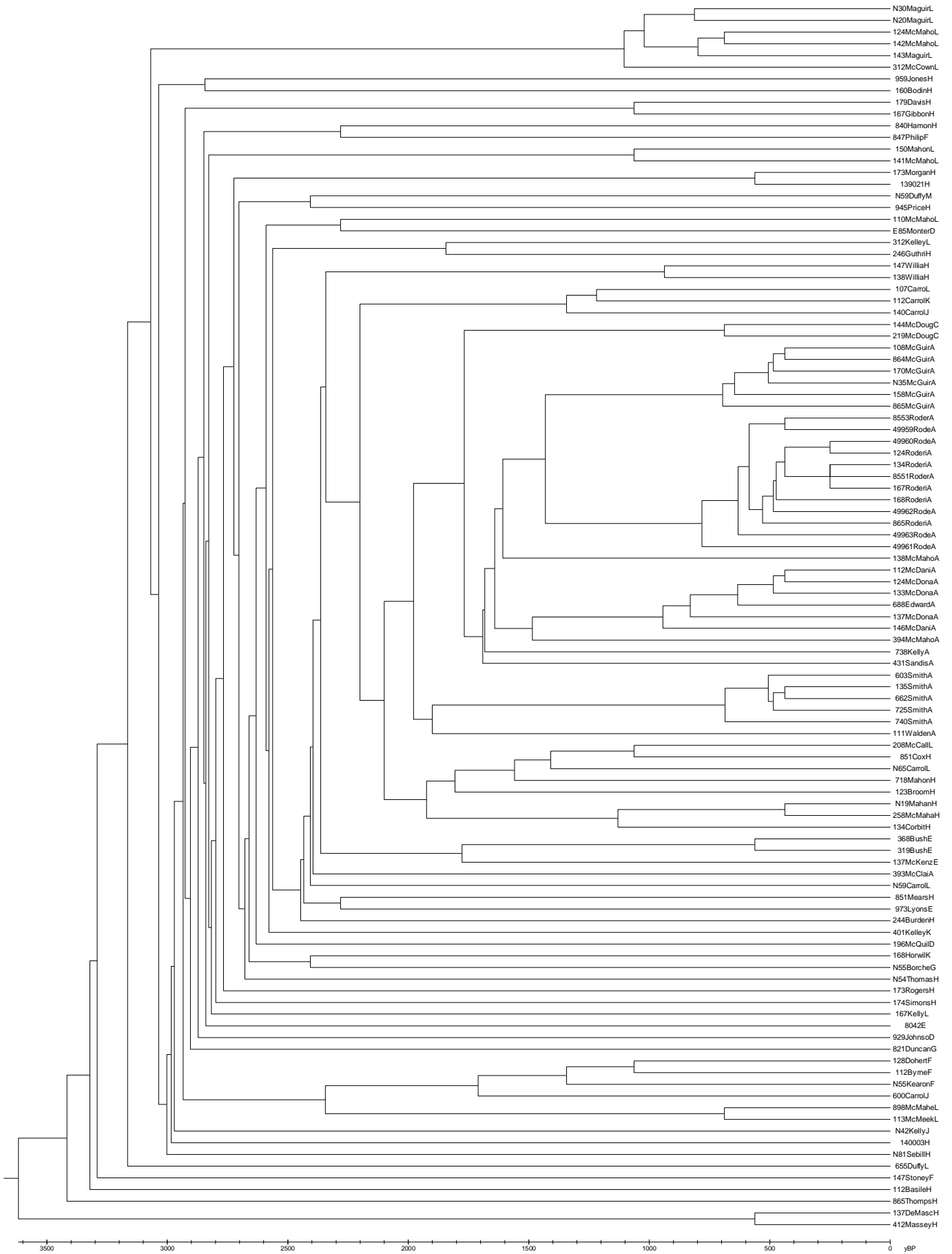


Figure B1. The ‘Outtree’ file from experiment 1 in ‘Input Order’ format with arbitrary time scale.

B.2. Experiment 2

The data entries for testers in column 2 of Table A 1 were processed through the Kitsch Application with 99 iterations. Associated symbols (Col 10, Table A1) were entered into the 'outtree' file for printing. Asterisks indicate those who were DF21+ (at time of data collection). The time scale is arbitrary and is as used in Experiment 1. The output in 'Input order' is given in Figure B.2.



Figure B2. The 'Outtree' file from experiment 2 in 'Input Order' format with arbitrary time scale

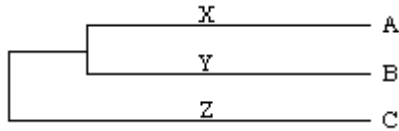
C. Construction of Phylogenetic Trees

The construction of phylogenetic trees using the Fitch & Margoliash methodology is described in this Annex, firstly by a worked example using actual data and secondly by presenting the tabular form of the output (Input Order Format) of the phylogenetic tree in Figure B2.

C.1. Fitch & Margoliash

Step 1:

Consider a simple tree:



One can estimate the distance between taxa A and its immediate ancestor X:

$$X = (D_{AB} + D_{AC} - D_{BC}) / 2$$

Similarly:

$$Y = (D_{AB} + D_{BC} - D_{AC}) / 2$$

$$Z = (D_{AC} + D_{BC} - D_{AB}) / 2$$

Step 2:

When there are more than three taxa, the third OTU is an average or composite of all distances except for the first two (A and B). The distance D_{AC} is the average of the distance between A and all the OTUs in C, and the distance D_{BC} is the average of the distance between B and all the OTUs in C.

Using the following data (tMRCA in years @ 99%p):

	169McDaniel	401McDaniel	N18Alexander	684McDaniel	150Calkin
169McDaniel	0	720	720	990	1410
401McDaniel		0	840	990	1410
N18Alexander			0	1260	1530
684McDaniel				0	1680
150Calkin					0

Where A is 169McDaniel and B is 401McDaniel, we get:

$$D_{AB} = 720$$

$$D_{AC} = (720 + 990 + 1410) / 3 = 1040$$

$$D_{BC} = (840 + 990 + 1410) / 3 = 1080$$

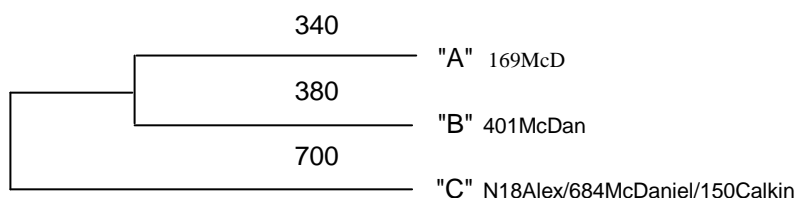
Then:

$$X = (720 + 1040 - 1080) / 2 = 340$$

$$Y = (720 + 1080 - 1040) / 2 = 380$$

$$Z = (1040 + 1080 - 720) / 2 = 700$$

Here, X is the Branch Length "A" and Y is the Branch Length "B", whereas Z is the distance from the composite OTU "C" to the Internal Node between 169McDaniel and 401McDaniel. This yields:



Step 3:

A and B are now combined as an OTU A/B and becomes the new "A"; in this case, this is 169McDaniel and 401McDaniel (McD/McD). The OTU that is now closest to AB becomes the new "B"; here it is N18Alexander. The remaining OTUs become the new composite "C" (684McDaniel and 150Calkin). Step 2 above is then repeated. The Distance from McD/McD to N18Alexander is the average: $(720 + 840) / 2 = 780$, and similarly for the distances from McD/McD to 684McDaniel and 150Calkin.

Thus:

	169McD/401McDan	N18Alex	684McDaniel	150Calkin
169McD/401McDan ("A")	0	780	990	1410
N18Alex ("B")		0	1260	1530
684McDaniel ("C")			0	1680
150Calkin ("C")				0

Then:

$$D_{AB} = 780$$

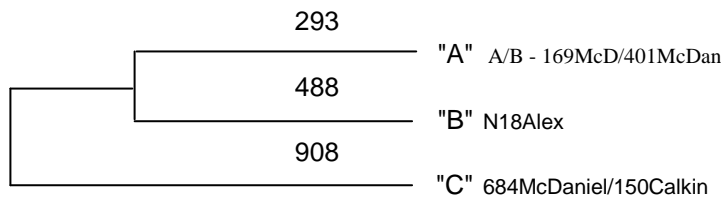
$$D_{AC} = (990 + 990 + 1410 + 1410) / 4 = 1200$$

$$D_{BC} = (1260 + 1530) / 2 = 1395$$

$$X = (780 + 1200 - 1395) / 2 = 293$$

$$Y = (780 + 1395 - 1200) / 2 = 488$$

$$Z = (1200 + 1395 - 780) / 2 = 908$$



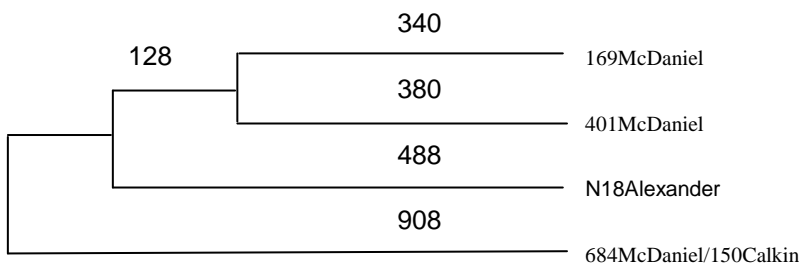
X is a composite of the 169McDaniel and 401McDaniel Branch Lengths of 340 and 380, or average of 360, and the Internal Branch length IB.

$$\text{Thus: } D_{AB} = 780 = 360 + \text{IB} + 293$$

$$\text{Thus: } \text{IB} = 780 - 360 - 293 = 127$$

$$\text{Alternatively: } \text{IB} = 488 - 360 = 128$$

This yields:



Step 4:

Step 3 is now repeated. 169McDaniel, N18Alexander and 401 McDaniel become the new "A", 684McDaniel becomes the new "B", and there is a single OTU remaining for "C", namely, 150Calkin.

The Distance from 169McDaniel, N18Alexander and 401 McDaniel to 684McDaniel is the average: $(990 + 990 + 1260) / 3 = 1080$... and similarly for 150Calkin: $(1410 + 1410 + 1530) / 3 = 1450$.

Thus:

	169McD/N18Alex/401McD	684McDaniel	150Calkin
169McD/N18Alex/401McD ("A")	0	1080	1450
684McDaniel ("B")		0	1680
150Calkin ("C")			0

Then:

$$D_{AB} = 1080$$

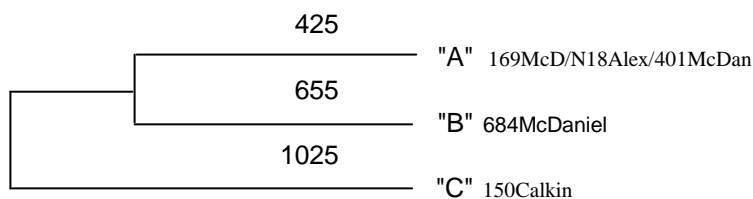
$$D_{AC} = (1410 + 1410 + 1530) / 3 = 1450$$

$$D_{BC} = 1680$$

$$X = (1080 + 1450 - 1680) / 2 = 425$$

$$Y = (1080 + 1680 - 1450) / 2 = 655$$

$$Z = (1450 + 1680 - 1080) / 2 = 1025$$

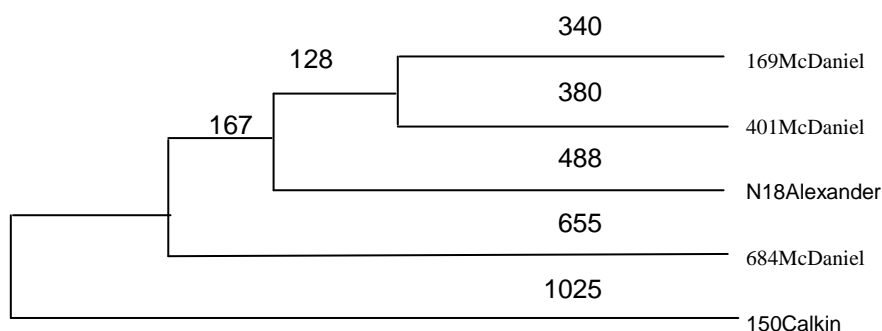


X is a composite of 169McDaniel, 401 McDaniel and N18Alexander, and Internal Branch Lengths of 488, 380, 340, and 128, respectively, plus a new Internal Branch Length. The average of the Branch Lengths is given by: $[(340 + 128) + (380 + 128) + 488] / 3 = 488$ the new Internal Branch Length IB then becomes: $655 - 488 = 167$ alternatively:

$$DAB = 1080 = IB + 488 + 655$$

$$\text{Thus: } IB = 1080 - 488 - 655 = 167$$

This yields a Fitch-Margoliash Unrooted Tree:

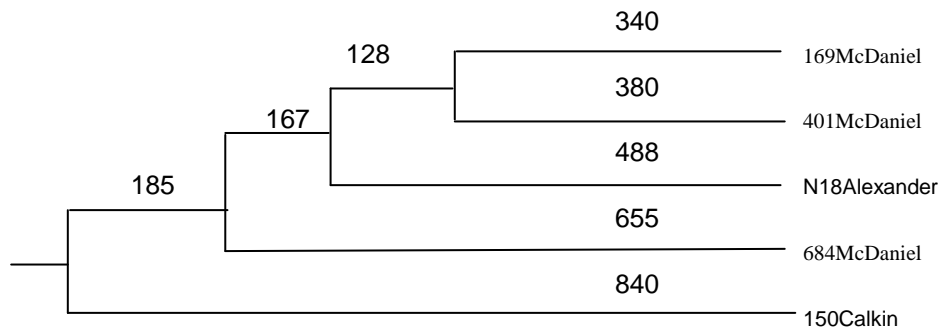


In the absence of additional information, one can assume that the Branch Length to, for example, 684McDaniel and to 150Calkin are close to being the same (assumes identical Molecular Clocks). This is similar to what is done in UPGMA. Then the new Internal Branch Length IB is given by:

$$1025 - 2 * IB = 655$$

$$\text{Thus: } IB = (1025 - 655) / 2 = 185$$

with a Branch Length from the Root to 150Calkin of: $1025 - 185 = 840$



Thus, Fitch-Margoliash:

- a. compares Taxa Distances in groups of three using the above Distance relationships
- b. uses composite OTUs in doing this which are averages of all OTUs present.

C.2. Tabular form of Phylogenetic Tree

The output data (outtree) used to generate the phylogenetic tree in Figure B2 is presented here in tabular form. The process within the Kitsch Programme is presumed to broadly follow the worked example given in Annex C.1. Ancestor IDs are numbers attributed by the software during the iteration processes as described in Annex C.1. Branch lengths are in generations (to convert to years, multiply by 30).

Ancestor ID	Descendent 1	Descendent 2	Branch Length 1	Branch Length 2
124	128Doherty	112Byrne	17.0	17.0
125	124	623Carroll	8.5	25.5
126	125	600Carroll	1.9	27.4
127	126	898McMahen	11.8	39.2
128	127	141McMahon	3.5	42.7
129	312Kelley	756Harris*	27.5	27.5
130	129	137McKenzi	6.5	34.0
131	425Roach	185Carrol*	27.5	27.5
132	122OMoore*	155Carragh	25.5	25.5
133	132	718Mahon	2.0	27.5
134	133	258McMahan	3.6	31.1
135	145Griffe*	N41Mahan	27.5	27.5
136	N65Carroll	208McCall	19.5	19.5
137	648White	219McDoug*	19.5	19.5
138	278Shannon	648Higgins	17.0	17.0
139	169Mcconne	786Conley	19.5	19.5
140	738Kelly	N47Burns	15.0	15.0
141	140	107Kelly	4.2	19.2
142	160Kelly	209McGuire	19.5	19.5
143	145McMaho*	799Hughes*	13.0	13.0
144	143	314Carrol	2.9	15.9
145	720Neal	372McInns*	13.0	13.0
146	145	641McKenna	0.9	13.9
147	146	N79McAtee	1.7	15.6
148	147	N88Connol*	2.8	18.4
149	148	117Clark	1.3	19.7

150	144	149	4.6	0.7
151	621Judd	166Hart*	15.0	15.0
152	N57OToole*	127Biggin*	11.0	11.0
153	152	209Plunkit	2.9	13.9
154	N49Heaney	401McDani*	11.0	11.0
155	154	829McQuil*	2.9	13.9
156	155	862Boylan*	1.7	15.6
157	156	368Paden*	0.3	15.9
158	189Main	100Duffy*	11.0	11.0
159	195McDona*	514Devine	13.0	13.0
160	159	334Carroll	0.9	13.9
161	171McKenna	123Clarke*	15.0	15.0
162	160	161	1.7	0.6
163	158	162	4.9	0.3
164	163	182OGuin*	1.2	17.1
165	157	164	1.7	0.5
166	165	N16Calkin*	1.2	18.8
167	153	166	5.1	0.1
168	167	477McAula*	0.4	19.4
169	145Robert*	N41Carrol*	17.0	17.0
170	168	169	0.4	2.8
171	170	191Cain*	0.1	19.9
172	782Shea	171	20.3	0.3
173	180Lawler	135Lynch	19.5	19.5
174	172	173	0.3	1.1
175	151	174	6.1	0.5
176	150	175	0.9	0.3
177	176	394McMahon	0.6	21.9
178	855Roderic	218Rice	19.5	19.5
179	177	178	0.6	3.0
180	142	179	3.2	0.2
181	141	180	3.6	0.1
182	181	N35McGuire	0.6	23.4
183	182	210ONeill*	0.1	23.5
184	183	137McDona*	0.1	23.6
185	184	982Feehan	0.1	23.7
186	185	N50Boyle	0.4	24.1
187	186	794Martin*	0.1	24.3
188	187	849Kelly	0.2	24.5
189	139	188	5.3	0.3
190	138	189	7.9	0.2
191	190	231McGuir*	0.3	25.3
192	191	N52Rogers*	0.4	25.7
193	192	138McMahon	0.1	25.8
194	193	N68Newell	0.4	26.1
195	194	971ODonogh	0.4	26.5

196	195	194McKenna	0.2	26.7
197	196	118Goodwin	0.4	27.1
198	137	197	7.8	0.1
199	198	DWXMcdonal	0.4	27.7
200	199	740Smith	0.5	28.2
201	136	200	10.2	1.4
202	201	156L21proj	0.2	29.8
203	202	704Clark	0.3	30.1
204	203	125Conley	1.1	31.1
205	204	179McCain	0.1	31.2
206	135	205	4.7	1.1
207	134	206	1.3	0.2
208	207	209Coleman	0.8	33.2
209	131	208	6.6	0.9
210	209	393McClai*	0.3	34.4
211	210	275Cooper	0.4	34.8
212	211	138William	1.6	36.4
213	212	244Burden	1.4	37.8
214	130	213	3.9	0.1
215	110McMahon	N59Carroll	29.5	29.5
216	215	162Moore	5.2	34.7
217	216	401Kelley	3.1	37.8
218	214	217	1.4	1.5
219	218	196McQuilk	0.4	39.7
220	219	761Caine*	1.1	40.7
221	220	N59Duffy	0.2	41.0
222	221	246Guthri	0.3	41.3
223	222	N54Thomas	0.2	41.4
224	223	E85Montero	0.8	42.2
225	224	204Hughes	0.4	42.6
226	945Price	130Duffy	38.5	38.5
227	225	226	0.1	4.2
228	173Morgan	174Simons	41.0	41.0
229	227	228	0.2	1.9
230	229	173Rogers	0.9	43.8
231	230	167Kelly	0.4	44.1
232	231	821Duncan	1.0	45.1
233	232	929Johnson	0.2	45.3
234	128	233	3.1	0.4
235	N20Maguire	142McMahon	17.0	17.0
236	235	312McCown	1.7	18.7
237	236	755Morgan#	23.1	41.8
238	234	237	0.9	4.8
239	N42Kelly	238	46.8	0.1
240	239	179Davis	0.1	46.9
241	240	959Jones	0.9	47.8

242	437Connel#	655Duffy	45.5	45.5
243	241	242	1.6	3.9
244	243	865Thompso	5.9	55.3
245	244	137DeMasci	2.0	57.2

D. Veracity of Phylogeny

The results presented in this analysis are dependent on the inferences made by the software package (Kitsch) from the tMRCA data generated in the McGee Utility as to which lineages were nearest neighbours and where they branched from each other in the evolutionary tree. As the method operates on essentially genetic distances (after conversion to tMRCA values) and is independent of modal values and names, two checks were carried out to test its validity.

ID	Genetic Distance																												
	D	1	N	3	1	1	6	6	8	6	3	1	5	N	2	N	1	7	1	N	2	3	1	1	3	1	2	2	5
DURRQ	-	22	20	26	21	22	25	22	4	3	4	2	2	4	12	13	12	14	6	8	5	3	4	11	11	11	11	5	5
142McMahon	22	-	6	6	21	22	23	24	21	23	19	24	24	21	14	21	20	20	24	24	25	23	26	19	19	19	19	22	22
N20Maguire	20	6	-	8	22	23	24	25	20	21	17	22	22	20	16	21	19	20	26	24	23	23	24	19	19	21	19	22	22
312McCown	26	6	8	-	23	22	24	24	22	27	23	28	28	25	16	21	21	22	26	26	27	27	30	23	23	21	21	26	25
112Byrne	21	21	22	23	-	6	12	12	23	22	19	23	23	20	20	21	24	16	25	23	25	22	23	23	23	25	23	21	20
128Doherty	22	22	23	22	6	-	12	9	23	23	20	24	24	21	21	22	22	17	24	22	24	23	26	24	24	24	22	22	21
623Carroll	25	23	24	24	12	12	-	14	26	26	23	27	27	24	22	24	24	20	27	25	25	26	29	25	25	25	23	25	22
600Carroll	22	24	25	24	12	9	14	-	23	23	20	24	24	21	19	20	19	19	24	22	24	23	26	24	24	22	22	20	21
829McQuill	4	21	20	22	23	23	26	23	-	5	6	6	6	6	10	12	11	13	9	9	8	7	8	13	13	12	12	7	7
652Boylan	3	23	21	27	22	23	26	23	5	-	5	5	5	5	13	14	13	15	9	9	8	6	7	12	12	12	12	6	6
368Paden	4	19	17	23	19	20	23	20	6	5	-	6	6	6	11	12	13	13	10	8	8	7	8	11	11	11	11	7	7
189Main	2	24	22	28	23	24	27	24	6	5	6	-	4	6	14	15	14	16	8	10	7	5	6	13	13	13	13	7	7
514Devine	2	24	22	28	23	24	27	24	6	5	6	4	-	6	14	15	14	16	8	10	7	5	6	13	13	13	13	7	7
N49Heaney	4	21	20	25	20	21	24	21	6	5	6	6	6	-	13	13	11	16	8	8	7	7	6	11	11	11	11	7	7
208McCall	12	14	16	16	20	21	22	19	10	13	11	14	14	13	-	7	9	9	14	12	13	13	16	11	11	9	9	13	13
N65Carroll	13	21	21	21	21	22	24	20	12	14	12	15	15	13	7	-	10	11	13	11	12	14	15	14	14	12	12	14	13
155Carragh	12	20	19	21	24	22	24	19	11	13	13	14	14	11	9	10	-	14	14	14	11	15	14	15	15	13	13	12	13
718Mahon	14	20	20	22	16	17	20	19	13	15	13	16	16	16	9	11	14	-	16	14	15	15	18	18	18	14	16	13	14
127Biggins	6	24	26	26	25	24	27	24	9	9	10	8	8	8	14	13	14	16	-	6	5	5	6	13	13	9	11	9	9
N57OToole	8	24	24	26	23	22	25	22	9	9	8	10	10	8	12	11	14	14	6	-	7	7	8	11	11	9	9	9	9
209Plunkit	5	25	23	27	25	24	25	24	8	8	8	7	7	7	13	12	11	15	5	7	-	6	5	14	14	12	12	10	8
334Carroll	3	23	23	27	22	23	26	23	7	6	7	5	5	7	13	14	15	15	5	7	6	-	5	12	12	12	12	6	6
129McDonal	4	26	24	30	23	26	29	26	8	7	8	6	6	6	16	15	14	18	6	8	5	5	-	13	13	13	13	9	9
112Carroll	11	19	19	23	23	24	25	24	13	12	11	13	13	11	11	14	15	18	13	11	14	12	13	-	4	6	4	12	12
304Bowes	11	19	19	23	23	24	25	24	13	12	11	13	13	11	11	14	15	18	13	11	14	12	13	4	-	6	4	12	12
116Purcell	11	19	21	21	25	24	25	22	12	12	11	13	13	11	9	12	13	14	9	9	12	12	13	6	6	-	4	12	12
233Springer	11	19	19	21	23	22	23	22	12	12	11	13	13	11	9	12	13	16	11	9	12	12	13	4	4	4	-	12	12
271Hughes	5	22	22	26	21	22	25	20	7	6	7	7	7	7	13	14	12	13	9	9	10	6	9	12	12	12	12	-	5
569Calkins	5	22	22	25	20	21	22	21	7	6	7	7	7	7	13	13	13	14	9	9	8	6	9	12	12	12	12	5	-

Figure D1. Genetic distances for a Selection of Phylogeny Groupings.

1) A Genetic Distance matrix was generated (in McGee) for a data sample of null and non-null lineages which were grouped by the Phylogeny software. The results are shown in Figure D1 consisting of groupings, 142McMahon to 312 McCown, 112Byrne to 600Carroll, 829McQuillan to N49Heaney, 208 McCall to 718Mahon, 127Biggins to 129McDonald, 112Carroll to 233Springer and 271Hughes to 569Calkins. These results show clear clustering for all groupings, some more significant than others. Clustering from low GD (to DURRQ) groupings showed cross relationships as well which is explained in section 5.

2) An examination of the DYS marker values for the first two groupings in Figure D1 showed permutations of marker values which could be defining. These were not the only marker differences but those showing commonality and slight differences commensurate with the observed branching. Ancestors 126 and 236 had a common ancestor (238) about 1400 years earlier (c.3200 yBP).

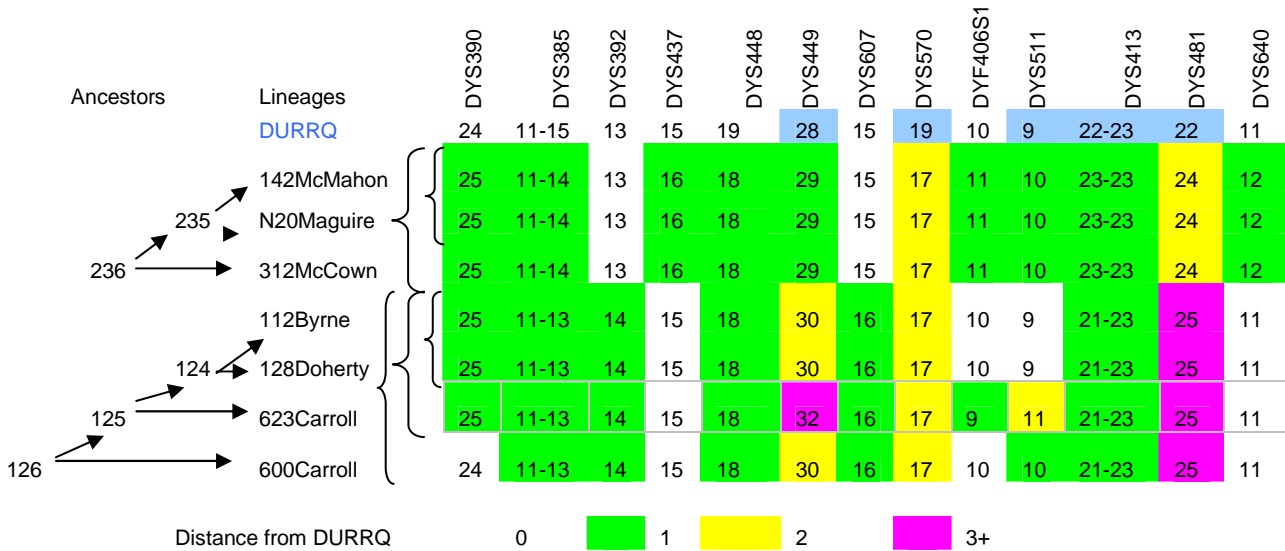


Table D1. Permutations of Marker Values and the Observed Branching in the Phylogeny Output.

Taken together, these results show a high degree of correlation between phylogeny produced results (where individual values cannot influence outcome) and clustering in McGee as well as patterns of marker values. In conclusion, it can be stated that Phylogeny as a method of identifying evolutionary nearest neighbours and their branching, is a sound methodology.

E. Calculating genetic distance for Y-DNA STR markers.

(Genebase Tutorials. Retrieved January 30, 2012, from <http://www.genebase.com/learning/article/46>)

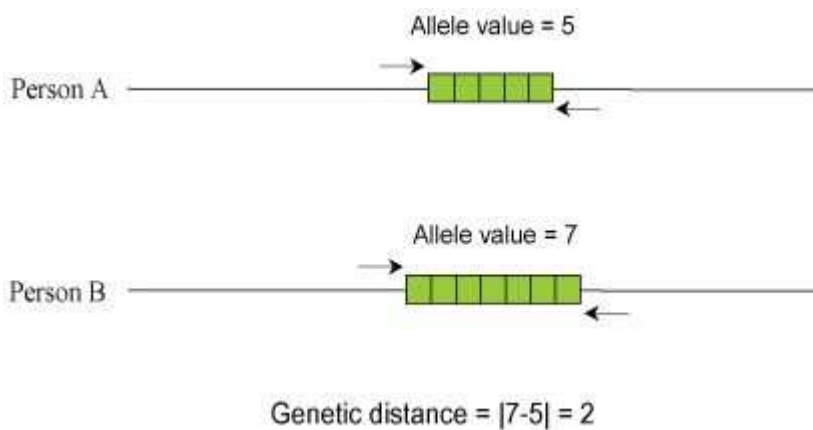
After you test your Y-DNA STR markers, you can compare your markers against the markers of any other male individual to see whether you share a common male ancestor. If a match is found, you can determine the tMRCA (time to most recent common ancestor), a measurement of approximately how long ago you and the matching individual shared the common ancestor.

A key measurement value when comparing the Y-DNA STR markers between two different individuals is **Genetic Distance**.

Genetic distance is a measurement of the total difference in allele values of different genetic markers between two individuals. The smaller the value of the genetic distance for a given set of STR markers, the closer two individuals are related, and the more recently they shared a common ancestor (tMRCA). The method used to determine genetic distance for four different Y-DNA STR marker types is explained below.

A. Calculation of genetic distance for single-copy Y-DNA STR markers

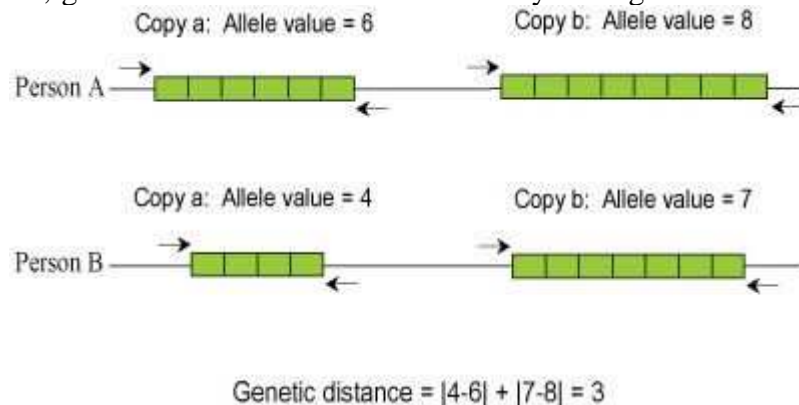
For single-copy STR markers, the calculation is straightforward. The genetic distance for each single copy marker between two individuals is the absolute value of the difference between the values of the markers:



The total genetic distance between two individuals is the sum of the genetic distances of all markers compared.

B. Calculation of genetic distance for multi-copy markers Markers DYS385, DYS459, DYS464 and YCAII are multi-copy Y-STR markers.

For most multi-copy markers, genetic distance can be calculated by adding the differences in allele values

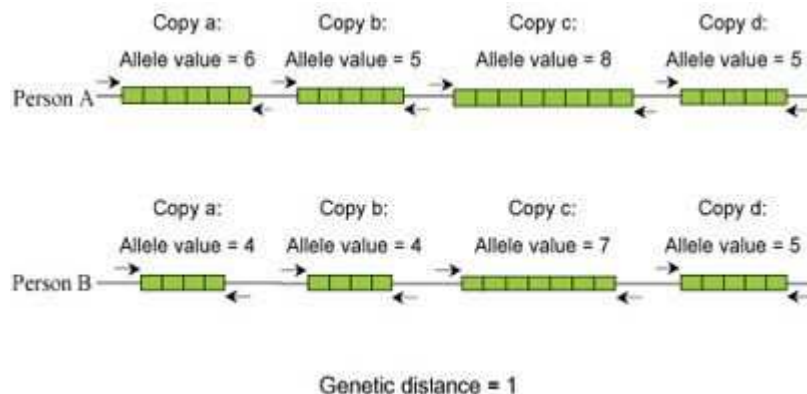


for each of the two copies.

C. Calculation of genetic distance for multi-copy marker DYS464 - using Infinite allele model

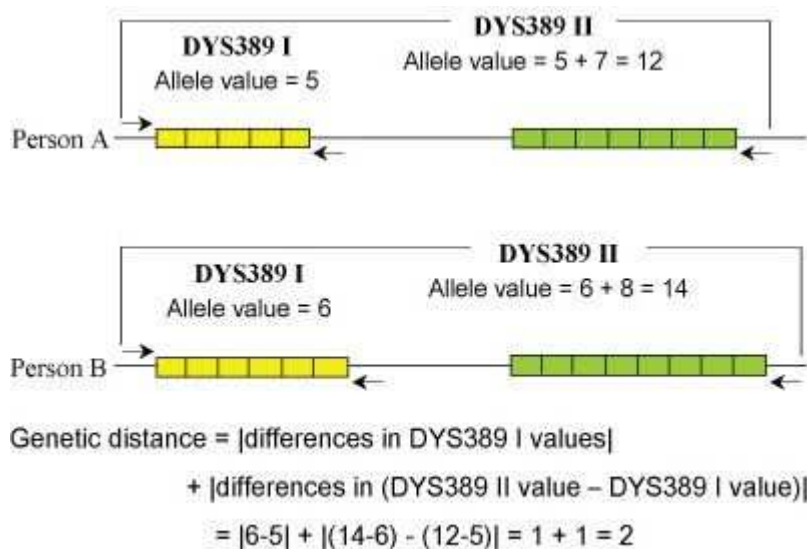
Assuming mutations at different copies of the same marker took place in a single generation, the Infinite allele method counts the total difference between all copies of the same marker as 1, despite the fact that more than one mutation exists.

The genetic distance for DYS464 is calculated using this method.



D. Calculation of genetic distance for DYS389i/ii

DYS389i is embedded in DYS389ii; therefore, the DYS389i values are included in DYS389ii values. Genetic distance can be determined by adding up two differences: differences in DYS389i values and differences in the second part of DYS389ii values, which are obtained by subtracting the DYS389ii values by DYS389i values.



The DNA Reunion feature will automatically calculate the Genetic Distance and tMRCA values between you and your matches. You can also perform your own matches by plugging in the marker values at www.dnacalculator.org.

F. Glossary

Allele. An allele is an alternative form of a gene (one member of a pair) that is located at a specific position on a specific [chromosome](#).

Britain. For the purposes of this analysis, Britain refers to mainland Britain comprising England, Scotland and Wales. The British Isles comprise mainland Britain and Ireland.

Celts. One of an Indo-European people originally of central Europe and spreading to Western Europe, the British Isles, and southeast to Galatia during pre-Roman times, especially to Briton and Gaul.

Clades. The clade is a hypothetical construct based on experimental data. Clades are found using multiple (sometimes hundreds) of traits from a number of species (or specimens) and analysing them statistically to find the most likely phylogenetic tree for the group.

Clan Colla. In the beginning of the 4th century, three warlike princes, called the [Three Collas](#), sons of Eochy Doimhlein, son of [Cairbre Lifeachar](#), legendary High King of Ireland, of the race of [Eremon](#), made a conquest of a great part of Ulster, which they wrested from the old possessors, princes of the race of Ir, called the Clanna Rory, or Rudericians. The names of the three chiefs were [Colla Uais](#), or Colla the noble, [Colla Meann](#), or Colla the famous, and [Colla da Chroich](#), or Colla of the two territories. [Colla Uais](#) became monarch of Ireland 327, and died in 332. The territory conquered by the three Collas comprised the present counties of [Louth](#), [Monaghan](#), and [Armagh](#).

Cluster. A group of testers whose DYS markers are sufficiently similar to indicate significant relationship to each other.

Dal Riada. **Dál Riata** (also **Dalriada** or **Dalriata**) was a [Gaelic](#) over kingdom on the western coast of [Scotland](#) with some territory on the northeast coast of [Ireland](#). In the late 6th and early 7th century it encompassed roughly what is now [Argyll and Bute](#) and [Lochaber](#) in Scotland and also [County Antrim](#) in Ireland.

DNA. Deoxyribonucleic acid is a [nucleic acid](#) containing the [genetic](#) instructions used in the development and functioning of all known living [organisms](#) (with the exception of [RNA viruses](#)).

DURRQ. The ID of the Colla modal. This is a 67 marker profile consisting of the most usual marker values for the 67 markers derived from a large group of Colla testers.

DYS. **DYS** is jargon and thought to be short for **DNA-Y** chromosome **STR**.

DYS 425. A specific marker whose value is the main determinant in distinguishing Colla from non-Colla DNA.

Evolutionary Tree. An evolutionary tree or phylogenetic tree is a branching diagram showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics. The taxa joined together in the tree are implied to have descended from a common ancestor. In a rooted phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of the descendants, and the edge lengths in some trees may be interpreted as time estimates.

Excel. Microsoft Excel spreadsheet.

FTDNA Family Tree DNA – the testing company used by most testers.

GD. **Genetic distance** is a measurement of the total difference in allele values of different genetic markers between two individuals. (See Annex E for a full explanation).

Haplogroup. In [molecular evolution](#), a **haplogroup** is a group of similar [haplotypes](#) that share a [common ancestor](#) having the same [single nucleotide polymorphism](#) (SNP) mutation in both haplotypes.

Haplotype. A **haplotype** in [genetics](#) is a combination of [alleles](#) (DNA sequences) at adjacent locations ([loci](#)) on the [chromosome](#) that are transmitted together.

Kitsch Programme. The Kitsch program of PHYLIP was used to calculate phylogram tree files.

ID The numerical or alpha numerical identity assigned by the testing company to a DNA sample.

Lineage. Lineage (anthropology) or [kinship](#), descent group that can demonstrate their common descent from an apical ancestor or a direct line of descent from an ancestor.

Marker. A **genetic marker** is a gene or [DNA sequence](#) with a known location on a [chromosome](#) that can be used to identify individuals or [species](#). It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed. A genetic marker may be a short DNA sequence, such as a sequence surrounding a single base-pair change ([single nucleotide polymorphism](#), SNP), or a long one, like [minisatellites](#).

McGee Utility. Y-DNA Comparison Utility

Modal. The mode is the number which appears most often in a dataset.

Mutation. A Mutation occurs when a DNA gene is damaged or changed in such a way as to alter the genetic message carried by that gene.

Null mutation. A null allele is a mutant [copy of a gene](#) that completely lacks that gene's normal function. At the [phenotypic](#) level, a null allele is indistinguishable from a deletion of the entire [locus](#).

Oriel. Oriel territory takes in the modern Counties Monaghan, Armagh and part of Louth.

PHYLIP. A Phylogeny Inference Package.

Phylogenetic tree See **Evolutionary Tree**.

Phylogram. A phylogram is a phylogenetic tree that has branch lengths proportional to the amount of character change.

Profile. The values (number of repeats) for STR markers for a given individual (67 markers in the current study).

R-L21+. Shorthand for haplogroup R1b1a2a1a1b4; the '4' indicates the fourth SNP downstream from P312 (in FTDNA nomenclature). This is a major European Clade thought to have occurred north of the Alps about 4,135 yBP.

R-DF21+. R-DF21 is a subclade under R-L21--two different SNPs. R-DF21 is comparatively quite old, with a tMRCA of 3112 yBP versus 4135 yBP for all of R-L21.

Singleton. A profile without any close matching to other profiles.

SNP. [Single Nucleotide Polymorphism](#), (SNP) is a single base-pair change.

STR. A **Y-STR** is a [short tandem repeat](#) (STR) on the [Y-chromosome](#). Y-STRs are often used in [forensics](#), paternity, and [genealogical DNA testing](#).

Taxon A taxonomic category or group, such as a phylum, order, family, genus, or species. Animal or plant group having natural relations.

Testers. Those who have had their DNA analysed for Y-chromosome markers.

tMRCA. Time to Most Recent Common Ancestor.

Y-DNA. Y-DNA is passed solely along the [patrilineal](#) line, from father to son. It does not recombine and thus Y-DNA changes only by chance mutation at each generation with no intermixture between parents' genetic material.

yBP. Years Before Present with the present being taken as A.D. 1950.